

# Computing the Observed Information Matrix for Dynamic Mixture Models

Michael J. Walsh  
Combat Systems Department



**Naval Undersea Warfare Center Division  
Newport, Rhode Island**

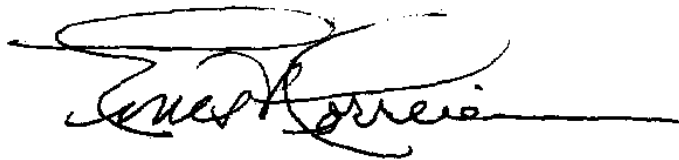
Approved for public release; distribution is unlimited.

## **PREFACE**

The work described in this report was funded by the Naval Undersea Warfare Center Division Newport's In-House Laboratory Independent Research (ILIR) Program.

The technical reviewer for this report was Marcus L. Graham (Code 2501).

**Reviewed and Approved: 25 September 2006**

A handwritten signature in black ink, appearing to read "Ernest Correia", with a long horizontal line extending to the right.

**Ernest Correia  
Head (acting), Combat Systems Department**



**REPORT DOCUMENTATION PAGE***Form Approved*  
**OMB No. 0704-0188**

Public reporting for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

**1. AGENCY USE ONLY (Leave blank)****2. REPORT DATE**

25 September 2006

**3. REPORT TYPE AND DATES COVERED****4. TITLE AND SUBTITLE**

Computing the Observed Information Matrix for Dynamic Mixture Models

**5. FUNDING NUMBERS****6. AUTHOR(S)**

Michael J. Walsh

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**Naval Undersea Warfare Center Division  
1176 Howell Street  
Newport, RI 02841-1708**8. PERFORMING ORGANIZATION  
REPORT NUMBER**

TR 11,768

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**Naval Undersea Warfare Center Division  
1176 Howell Street  
Newport, RI 02841-1708**10. SPONSORING/MONITORING  
AGENCY REPORT NUMBER****11. SUPPLEMENTARY NOTES****12a. DISTRIBUTION/AVAILABILITY STATEMENT**

Approved for public release; distribution is unlimited.

**12b. DISTRIBUTION CODE****13. ABSTRACT (Maximum 200 words)**

The observed information matrix for an important class of finite mixture models, called dynamic mixture models, is derived in this report. Dynamic mixture models are useful probability models for random data originating from a number of distinct moving sources. The multiple-target tracking problem is one application of these models. For these models, the inverse of the observed information matrix is a consistent estimate of the error-covariance matrix for the mixture parameters.

Measurement-to-source assignment uncertainty is unavoidable in these problems, and increases as the distance between sources in the sample space decreases. The observed information matrix computations presented here account for this uncertainty by subtracting the information in the unobserved assignments, treated as missing data, from the information in the expected complete data sample. Two target tracking examples are given that demonstrate these computations for the linear Gauss-Markov mixture model for multiple target tracking. In each case, the consistency of the resulting error-covariance matrices is examined.

**14. SUBJECT TERMS**

Dynamic Mixture Model; Expectation-Maximization; Finite Mixture Model; Gauss-Markov Process; Gaussian Mixture; Kalman Filter; Kalman Smoother; Multiple Target Tracking; Multivariate Linear Model; Observed Information Matrix; Probabilistic Multi-Hypothesis Tracking

**15. NUMBER OF PAGES**

88

**16. PRICE CODE****17. SECURITY CLASSIFICATION  
OF REPORT**

Unclassified

**18. SECURITY CLASSIFICATION  
OF THIS PAGE**

Unclassified

**19. SECURITY CLASSIFICATION  
OF ABSTRACT**

Unclassified

**20. LIMITATION OF ABSTRACT**

SAR



## TABLE OF CONTENTS

Section	Page
LIST OF ILLUSTRATIONS . . . . .	iii
LIST OF TABLES . . . . .	iv
1 INTRODUCTION . . . . .	1
1.1 Related Work . . . . .	2
1.2 Report Organization . . . . .	7
2 MAXIMUM LIKELIHOOD ESTIMATION USING THE EM METHOD . . . . .	9
2.1 General Case . . . . .	9
2.2 Independent Observations . . . . .	10
2.3 Maximum <i>A Posteriori</i> Estimation . . . . .	10
3 OBSERVED INFORMATION FOR INCOMPLETE DATA PROBLEMS . . . . .	13
3.1 General Case . . . . .	13
3.2 Independent Observations . . . . .	15
3.3 Posterior Observed Information . . . . .	17
4 OBSERVED INFORMATION FOR FINITE MIXTURE MODELS . . . . .	19
4.1 General Case . . . . .	19
4.2 Gaussian Mixtures . . . . .	20
5 OBSERVED INFORMATION FOR DYNAMIC MIXTURE MODELS . . . . .	25
5.1 Deterministic Motion . . . . .	25
5.2 Stochastic Motion . . . . .	31
6 THEORETICAL AND PRACTICAL CONSIDERATIONS . . . . .	41
6.1 Asymptotic Normality of $\hat{\theta}$ . . . . .	41
6.2 Sequential Versus Batch Processing . . . . .	41
7 EXAMPLES . . . . .	45
7.1 Estimator Consistency . . . . .	45
7.2 Two Crossing Targets . . . . .	48
7.3 Single Target in Clutter . . . . .	58

8	CONCLUSIONS . . . . .	65
8.1	Summary of Findings . . . . .	65
8.2	Alternative Approaches . . . . .	65
8.3	Future Investigations . . . . .	67
	APPENDIX A - APPROXIMATION TO THE OBSERVED INFORMATION MATRIX . . . . .	A-1
	APPENDIX B - INVERSE OF THE GAUSS-MARKOV PRIOR COVARIANCE MATRIX . . . . .	B-1
	APPENDIX C - ADDING A CLUTTER MODEL TO PMHT . . . . .	C-1
	REFERENCES . . . . .	R-1

## LIST OF ILLUSTRATIONS

Figure	Page
1 Distance Between Targets in $x$ -Dimension in Units of Measurement Standard Deviation for Crossing Targets Example . . . . .	50
2 Average NEES with 95% Acceptance Region for Crossing Targets Example with Batch Length 25 . . . . .	51
3 Average NEES with 95% Acceptance Region for Crossing Targets Example with Batch Length 10 . . . . .	51
4 Average NEES with 95% Acceptance Region for Crossing Targets Example with Batch Length 5 . . . . .	52
5 Average NEES with 95% Acceptance Region for Crossing Targets Example with Batch Length 1 . . . . .	52
6 K, CVM, and AD Statistics with 95% Acceptance Regions for Crossing Targets Example with Batch Length 25 . . . . .	53
7 K, CVM, and AD Statistics with 95% Acceptance Regions for Crossing Targets Example with Batch Length 10 . . . . .	54
8 K, CVM, and AD Statistics with 95% Acceptance Regions for Crossing Targets Example with Batch Length 5 . . . . .	55
9 K, CVM, and AD Statistics with 95% Acceptance Regions for Crossing Targets Example with Batch Length 1 . . . . .	56
10 Average NEES with 95% Acceptance Region for Single Target in Clutter Example with Batch Lengths 25 and 10 . . . . .	60
11 Average NEES with 95% Acceptance Region for Single Target in Clutter Example with Batch Lengths 5 and 1 . . . . .	60
12 K, CVM, and AD Statistics with 95% Acceptance Region for Single Target in Clutter Example with Batch Lengths 25 and 10 . . . . .	61
13 K, CVM, and AD Statistics with 95% Acceptance Region for Single Target in Clutter Example with Batch Lengths 5 and 1 . . . . .	62

## LIST OF TABLES

Table		Page
1	Percentage of NEES, K, CVM, and AD Values That Fall Within Their Respective 95% Acceptance Regions for the Crossing Targets Example . . . . .	57
2	Percentage of NEES, K, CVM, and AD Values That Fall Within Their Respective 95% Acceptance Regions for the Single Target in Clutter Example . .	63



# COMPUTING THE OBSERVED INFORMATION MATRIX FOR DYNAMIC MIXTURE MODELS

## 1. INTRODUCTION

Mixture distributions are widely used in statistical analysis to model data originating from a number of distinct sources. In these models, each source is represented by a component of the mixture. The distance between the sources in the sample space determines both the level of difficulty of estimating the mixture parameters, and the amount of uncertainty in the parameter estimates. Intuitively, the uncertainty associated with assigning measurements to sources should increase as the distance between the sources in the sample space decreases. This report is concerned with accurately assessing the estimation error in mixture estimation problems for which measurement-to-source assignment uncertainty is a major contributor to uncertainty in the data.

Mixture estimation can be treated as a missing data problem where the missing data are the measurement-to-source assignments. Consequently, the expectation-maximization (EM) method of Dempster et al. [1] provides a convenient iterative approach for finding maximum likelihood estimates of the mixture parameters. Indeed, mixture estimation is one of the numerous applications of the EM method discussed in their paper. The EM method is most useful in situations where the corresponding complete data (i.e., observed data + missing data) problem has a straightforward solution. Gaussian mixture estimation is one example. In this case, application of the EM method yields a sequence of weighted linear least-squares estimates for each of the Gaussian mean vectors that converge to their maximum likelihood estimates. Gaussian mixture models have been studied extensively by many authors (see, for example, the monograph by McLachlan and Basford [2] and the references therein) and are the basis for the more complex mixture models considered here.

The main criticisms of the EM method are that it does not give an immediate expression for the error-covariance matrix for the estimated parameters, and that it converges slowly near the solution (in contrast to gradient-based techniques, which at least approximate the error-covariance matrix for the parameter estimates, and often converge rapidly near the solution). Louis [3] addresses both criticisms in his paper on finding the observed information matrix when using the EM method for incomplete data problems. In his paper, Louis shows that the observed information matrix, defined as minus the second derivative of the observed (or incomplete) data log-likelihood function, evaluated at the maximum likelihood estimate, is obtained by straightforward manipulations of the complete data log-likelihood function. The inverse of this matrix can then be used as an estimate of the error-covariance matrix for the

estimated parameters, as suggested by Efron and Hinkley [4]. Furthermore, Louis shows that the observed information matrix can be used to accelerate convergence of the EM iterations.

In this report Louis’s results are applied to an important class of mixture models termed “dynamic mixture models.” A dynamic (or time-varying) mixture model constitutes a sequence of standard mixture distributions related in time by a process model. Dynamic mixture distributions are used to model data collected over time originating from a number of distinct *moving* sources. (Here, source motion refers to a change over time of any characteristic of the source—for example, location, orientation, and intensity.) If the sources are stationary, one can pool data collected over multiple sampling times and use a standard (static) mixture to describe the sample distribution. However, if the sources are non-stationary, one must account for source motion in the mixture to accurately model the distribution of the sample at each sampling time. A dynamic mixture model may be deterministic or stochastic. In the former case, a parametric motion model is used to describe the trajectory of each source in the mixture. In the latter case, the trajectory of each source is treated as a sequence of random variables whose mean evolves according to a deterministic motion model. In either case, the objective of this report is to compute the observed information matrix for the estimated mixture parameters, and to assess the quality of this matrix (or, more precisely, the inverse of this matrix) as a characterization of estimation error.

## 1.1 RELATED WORK

Work related specifically to computing the observed information matrix for dynamic mixture models—and, more generally, to assessing the impact of measurement-to-source assignment uncertainty on estimation error—is found in both the statistics and engineering literature. One paper from the statistics literature is particularly relevant. In [5], Meng and Rubin propose the supplemented EM (SEM) algorithm as an alternative approach for computing the error-covariance matrix for parameter estimates obtained using the EM method. Their approach, in contrast to Louis’s analytical approach, is half analytical and half numerical. In short, the SEM algorithm requires analytical differentiation to obtain the information matrix associated with the complete data, but uses numerical differentiation to compute the information matrix associated with the missing data. The difference between these two matrices is the observed information matrix, which is then inverted to obtain the error-covariance matrix. For problems where the algebraic analysis required by Louis’s procedure is tedious or intractable, the SEM algorithm is an attractive alternative.

Two other papers from the statistics literature are worth mentioning. Green [6] discusses the EM method in the context of maximum penalized likelihood estimation (mathematically equivalent to maximum *a posteriori* estimation), a problem for which he laments

the EM method has seen little use. In his paper, Green makes a simple modification to the EM algorithm for maximum penalized likelihood estimation to obtain the one-step-late (OSL) algorithm, which is often easier to compute and converges at least as quickly. Green’s paper is relevant to this discussion because the estimation problem for stochastic dynamic mixture models is a maximum *a posteriori* estimation problem for which the OSL algorithm may be potentially useful. In a later paper, Segal et al. [7] combine the results of Meng, Rubin, and Green to compute error-variances via the SEM algorithm for maximum penalized likelihood estimates obtained using the OSL algorithm. Their approach is directly applicable to computing the error-covariance matrix for stochastic dynamic mixture models. However, this report will show that the algebraic analysis required by Louis’s approach yields insightful expressions for the particular stochastic dynamic mixture model considered here—namely, the linear Gauss-Markov model.

In the engineering literature, and the target tracking literature in particular, related work falls into three overlapping categories: mixture models for multiple target tracking, information reduction factors for single target tracking in clutter, and minimum variance (Cramér-Rao) bound calculations for tracking performance prediction. The report by Streit and Luginbuhl [8] and the paper by Gauvrit et al. [9] are the primary references for the mixture model approach to multiple target tracking considered in this report. In this approach, each target is represented by a component (or possibly a collection of components) in a mixture model for the measurement distribution. By the very nature of this model, it is assumed that every measurement originates from all the targets; more precisely, each measurement is assigned to every target with a certain probability. This unorthodox tracking model is a contradistinction to the widely accepted multiple hypothesis tracking (MHT) model proposed by Reid [10], in which each measurement is assigned to one and only one target, or to clutter (background noise). While the MHT assignment model is perhaps more realistic, it leads to a set of track hypotheses that grows exponentially with the number of measurements. Consequently, MHT algorithms require sophisticated heuristics to manage hypothesis enumeration, which typically involves pruning and merging branches on the hypothesis tree. Alternatively, the mixture model algorithms have complexity that is roughly linear in the numbers of measurements and targets; thus, hypothesis management is not required for these algorithms. However, as succinctly put by Streit in [11], the price to be paid for the “heresy” of violating the one-measurement-per-target rule of multiple target tracking is a likelihood function that may be “riddled” with local maxima. Streit goes on in [11] to blend a mixture model with “limited enumeration” to address this problem.

The stochastic dynamic mixture model discussed in this report is precisely the mixture model used by Streit and Luginbuhl [8] and Gauvrit et al. [9] for multiple target tracking.

Streit and Luginbuhl call the approach “probabilistic multi-hypothesis tracking (PMHT).” Of these works, only the former considers computation of the error-covariance matrices for the track estimates. However, the analysis of Streit and Luginbuhl is incomplete in that the matrices identified in their report as the error-covariance matrices for PMHT do not account for the information lost to the missing data. Hence, what have become widely accepted as the error-covariance matrices for PMHT are incorrect and, worse, as will be shown in this report, are overly-optimistic. This report gives a precise statistical interpretation of these matrices and derives expressions for the correct error-covariance matrices for PMHT that account for the information lost to the missing data.

Two additional approaches related to PMHT must also be acknowledged. Avitzour’s work [12], a remarkably similar but independent antecedent to PMHT, is perhaps the first application of missing data and EM to multiple target tracking. The similarities and differences between the two approaches are discussed in [8]; notably, PMHT substitutes a stochastic (Markovian) motion model for the deterministic (polynomial) motion model of Avitzour’s approach. Also, Avitzour does not discuss computation of error-covariance matrices for track estimates. The multiple target tracking approach proposed by Molnar and Modestino [13] also uses missing data and EM, although their measurement-to-target assignment model is markedly different than that of Avitzour’s or PMHT’s. Nevertheless, Molnar and Modestino propose an approximation to the error-covariance matrix for the target state estimates that explicitly accounts for measurement-to-target assignment uncertainty. In their approximation, each measurement’s contribution to the total information content of all the measurements with respect to each target is scaled by the measurement-to-target assignment probability. Neither the development nor the quality of this approximation is discussed in [13].\*

The notion that measurement assignment uncertainty in tracking should increase the variance of the track estimates is not new. For example, in [14] Fortmann et al. analyze the effect of clutter on the update of the target state covariance matrix. Specifically, they consider a deterministic approximation to the stochastic matrix Riccati equation associated with the probabilistic data association (PDA) filter of Bar-Shalom and Tse [15]. (Recall that the matrix Riccati equation of Kalman filtering theory is a recursion for the update of the state covariance matrix.) This approximation, which replaces the random (data-dependent) quantities in the stochastic matrix Riccati equation with their expected values, leads to a modified matrix Riccati equation that looks like the standard equation, with the addition of a scalar factor in front of the Kalman gain term. This scalar factor, called the information reduction factor and denoted  $q_2$  in their paper, takes values between 0 and 1; these extremes correspond to total assignment uncertainty and no assignment uncertainty, respectively. A

---

\*Note that the critical term in this approximation (the term in brackets in [13, equation (48)]) is missing an inverse.

value of  $q_2 = 0$  eliminates the Kalman gain term, so that the updated state covariance matrix equals the predicted state covariance matrix; a value of  $q_2 = 1$  reduces the modified matrix Riccati equation to the standard equation, so that the updated state covariance matrix is equal to the minimum state covariance matrix. The information reduction factor  $q_2$  is a function of the probability of detection  $P_D$  and the probability of false alarm  $P_{FA}$ , with  $q_2 = 1$  when  $P_D = 1$  and  $P_{FA} = 0$ , and  $0 \leq q_2 < 1$  when  $P_D \leq 1$  and  $P_{FA} > 0$ . In short, the information reduction factor accounts for measurement assignment uncertainty due to missed detections and clutter. Computation of  $q_2$  is nontrivial, and is described by Gelfand et al. in [16].

The most commonly used baseline for judging target tracking performance is the Cramér-Rao lower bound (CRLB), that is, the minimum variance bound on estimation error. Recall that the CRLB is defined in terms of an average (expectation) over all possible values of the observed data. Hence, for any given tracking model, the CRLB can be used to predict tracking performance in the absence of measurements. The multiple target tracking problem complicates computation of the CRLB, since the measurement-to-target assignments are almost never observed. The approach then is to marginalize over the assignment hypotheses and compute the minimum variance bound based on the marginal distribution of measurements and target states. This is the approach presented by Daum in [17]. As described by Daum, computation of this marginal is impractical, as the number of association hypotheses is enormous even for small problems. To address this problem, Daum provides a family of lower bounds on the minimum variance bound, where each member of the family corresponds to a collection of association hypotheses that include the correct hypothesis. The lower bound corresponding to the set of all possible hypotheses corresponds to the minimum variance bound, whereas the lower bound corresponding to the subset that contains *only* the correct hypothesis corresponds to the trivial lower bound, which ignores measurement assignment uncertainty entirely. Thus, tighter bounds can be achieved at the expense of computational complexity by considering progressively larger sets of association hypotheses.

Further approximations to the CRLB when there is measurement assignment uncertainty have been computed by Jauffret and Bar-Shalom [18] and Kirubarajan and Bar-Shalom [19] for tracking a single target in clutter. Both works use the approach of Fortmann et al. [14] to show that the Fisher information matrix (FIM) for this problem is a scalar multiple of the FIM for the problem without clutter, where the scalar multiple is the information reduction factor  $q_2$  developed in [14]. It follows that the CRLB (inverse of the FIM) increases rapidly with the amount of assignment uncertainty and, in fact, grows without bound as  $q_2 \rightarrow 0$ . Subsequent papers by Willett and Bar-Shalom [20] and Niu et al. [21] extend these results by finding a set of sufficient conditions for the class of models for single target tracking in clutter whose CRLB takes this form.

Another set of references relevant to this report deal with CRLB computations for the mixture model approach to multiple target tracking. In [22], Perlovsky develops explicit CRLB expressions for the parameters of a normal (Gaussian) mixture model of the measurement distribution for tracking multiple targets in clutter. This work is based on his earlier paper [23] on computing the CRLB for normal mixtures. While the CRLB expressions developed in these papers are indeed explicit, they are written in terms of quantities (“class overlap” terms) that ultimately require numerical evaluation of multi-dimensional integrals (expectations over all values of the observed data), where there are as many integrals as there are observations, and each integral has dimension equal to that of a data point. Le Cadre et al. [24] address the same problem in their paper on computing the CRLB for the multiple target version of a classic subject in the tracking literature known as bearings-only target motion analysis. As in Perlovsky’s papers, the class overlap or “source interaction” terms induced in the CRLB expressions by the mixture model for measurement-to-target assignment lead to integrals that cannot be evaluated in closed form. Among the contributions in [24] are analytical approximations to these integrals based on series expansions. In [25], Hue et al. derive recursive formulas for computing the “posterior” CRLB for multiple target tracking assuming stochastic (Markovian) target motion and three different measurement-to-target assignment models—namely, the known assignment model, the one-measurement-per-target model, and the PMHT model. The posterior CRLB for the PMHT model is closely related to the posterior observed information matrix derived in this report; specifically, the posterior CRLB is the inverse of the expected value of the posterior information matrix over all values of the observed data and all values of the target states. These expectations are evaluated in Hue et al. [25] using Monte-Carlo integration techniques. Notably, Hue et al. show that measurement assignment uncertainty raises the posterior lower bound on estimation error, often substantially. Analogous results are obtained in this report with regard to the impact of measurement assignment uncertainty on the *in-situ* assessment of estimation error given by the inverse of the posterior observed information matrix.

The work most relevant to the present discussion among this set of references is the report by Graham and Streit [26], which discusses computation of the CRLB for PMHT and which shows that the Fisher information matrix for PMHT is equal to the Fisher information matrix derived from the complete data likelihood function, minus the information matrix associated with the missing data (measurement-to-target assignments). This result is a manifestation of the “missing information principle” to be discussed in more detail later in this report. Thus, the complete data lower bound obtained by inverting the complete data Fisher information matrix, which Graham and Streit show to be block-diagonal, is a lower bound on the CRLB. This lower bound on the lower bound, they argue, is analogous to the trivial



lower bound of Daum's approach. The present report is not concerned with the computation of the Fisher information matrix and the minimum variance bound, which are independent of observed data, but rather with the observed information matrix and its inverse, and their assessment of estimation accuracy as a function of the measurements. Furthermore, explicit closed-form expressions for the complete data and missing data information matrices, in terms of the maximum likelihood estimates for the mixture parameters, are derived.

Finally, a recent paper by Cai et al. [27] presents an EM algorithm for tracking maneuvering targets in clutter, and uses the SEM algorithm to compute the error-covariance matrix for the estimated target parameters. Theirs appears to be the first use of the SEM algorithm in the tracking literature. In fact, their algorithm is essentially Avitzour's algorithm with position and amplitude measurements. Hence their paper contains the first accurate computation in the literature of the error-covariance matrix for a PMHT-related algorithm. In contrast, the error-covariance matrix for PMHT is computed in this report using the analytical approach of Louis. The benefit of Louis's approach in this case is that it gives precise statistical meaning to certain quantities fundamental to the PMHT computations that have often been incorrectly interpreted as the error-covariance matrices for PMHT.

## 1.2 REPORT ORGANIZATION

To establish terms and notation used throughout this report, the EM method for maximum likelihood estimation for incomplete data problems is introduced in section 2. In section 3, Louis's derivation of the observed information matrix for the general case of incomplete data is summarized, and a simplification of his expressions for the special case of independent observations is presented. Also in this section, the posterior observed information matrix, which is used to compute the error-covariance matrix for stochastic dynamic mixture models, is defined. In section 4, maximum likelihood estimation for finite mixture models using the EM method is reviewed, and general expressions for the corresponding observed information matrix are presented. Expressions for the maximum likelihood estimates and the observed information matrix for Gaussian mixture models are also given in this section.

The observed information matrix and the posterior observed information matrix for the deterministic and stochastic dynamic mixture models, respectively, are derived in section 5. In both cases, the maximization step at the final EM iteration is shown to be equivalent to a standalone estimation problem for which the error-covariance matrix is given by the inverse of the complete data information matrix in the case of deterministic motion, and the inverse of the posterior complete data information matrix in the case of stochastic motion. The latter result provides a precise statistical interpretation of the error-covariance matrices obtained as byproducts of PMHT for the linear-Gaussian case.

Section 6 discusses the suitability of the inverse of the observed information and posterior observed information matrices as estimates of the error-covariance matrices for the deterministic and stochastic dynamic mixture models, respectively, and the cost of computing these inverses when the number of sampling times is large. Section 7 includes two target tracking examples using the stochastic dynamic mixture model, one of two crossing targets, and one of a single target in clutter. The consistency of the target parameter estimates is examined for each example. The report concludes in section 8 with a summary of findings, a discussion of alternative approaches for computing the observed information matrix, in particular the SEM algorithm, and a discussion of topics for future investigation.



## 2. MAXIMUM LIKELIHOOD ESTIMATION USING THE EM METHOD

### 2.1 GENERAL CASE

Consider the general incomplete data problem in which there are two sample spaces, the complete data sample space  $\mathcal{X}$  and the incomplete (or observed) data sample space  $\mathcal{Y}$ , and a many-to-one mapping  $Y : \mathcal{X} \rightarrow \mathcal{Y}$ . Let  $x$  denote an arbitrary point in  $\mathcal{X}$ . The point  $x$  is not observed directly; rather, the point  $y = Y(x)$  in  $\mathcal{Y}$  is observed. Assume a family of density functions  $f_X(x; \theta)$  on  $\mathcal{X}$  indexed by the parameter vector  $\theta$  from a space  $\Omega$ . Let  $L = \{x : Y(x) = y\}$  denote the section of  $\mathcal{X}$  determined by  $y$ . The corresponding family of observed data density functions  $f_Y(y; \theta)$  on  $\mathcal{Y}$  is given by

$$f_Y(y; \theta) = \int_L f_X(x; \theta) dx, \quad (2-1)$$

where integration here is meant in the most general sense.

For a fixed sample  $y$ , the density function  $f_Y(y; \theta)$  taken as a function of the parameter vector  $\theta$  is the incomplete or observed data likelihood function. Let  $\lambda_Y(y; \theta) = \log f_Y(y; \theta)$  denote the observed data log-likelihood (or support) function. The maximum likelihood estimate of the parameter vector  $\theta$ , denoted  $\hat{\theta}$ , is that value of  $\theta$  in the space  $\Omega$  that maximizes  $f_Y(y; \theta)$  or, equivalently,  $\lambda_Y(y; \theta)$  for the given sample  $y$ ; that is,

$$\hat{\theta} = \arg \max_{\theta} \lambda_Y(y; \theta). \quad (2-2)$$

Let  $\lambda_X(x; \theta)$  denote the complete data support function for a fixed sample  $x$ . Using the EM method, the maximum likelihood estimate of  $\theta$  is obtained by solving the following sequence of complete data problems:

$$\theta^{(k+1)} = \arg \max_{\theta} E_{\theta^{(k)}}[\lambda_X(X; \theta) \mid X \in L] \quad (2-3)$$

for  $k = 0, 1, \dots$ , where  $\theta^{(0)}$  is an initial estimate of  $\theta$ , and

$$E_{\theta^{(k)}}[\lambda_X(X; \theta) \mid X \in L] = \int_L \lambda_X(x; \theta) f_{X|Y}(x|y; \theta^{(k)}) dx \quad (2-4)$$

is the conditional expectation of the complete data support function at the  $k$ th iteration. Assuming that the observed data density functions  $f_Y(y; \theta)$  are strictly positive, the conditional density functions  $f_{X|Y}(x|y; \theta)$  are defined by

$$f_{X|Y}(x|y; \theta) = \frac{f_X(x; \theta)}{f_Y(y; \theta)} = \frac{f_X(x; \theta)}{\int_L f_X(x; \theta) dx}. \quad (2-5)$$

Expressions (2-4) and (2-3) are the expectation step, or E-step, and the maximization step, or M-step, respectively, of the EM method. The sequence  $\theta^{(k)}$  of EM iterates converges to the

maximum likelihood estimate  $\hat{\theta}$  under the regularity conditions stated in Dempster et al. [1] and Wu [28]. These conditions are assumed to hold here and throughout the report. The usual regularity conditions for the existence of maximum likelihood estimates and information matrices, and the interchangeability in order of differentiation and integration, are also assumed in the sequel. (See, for example, Cramér [29, chapters 7, 32, and 33] and Casella and Berger [30, chapters 2 and 7] for statements of these conditions.)

## 2.2 INDEPENDENT OBSERVATIONS

Suppose the complete data  $x$  consists of  $n$  independent (but not necessarily identically distributed) samples  $x_1, \dots, x_n$ . These samples are not observed directly; rather, the samples  $y_1 = Y_1(x_1), \dots, y_n = Y_n(x_n)$  through the many-to-one mappings  $Y_1 : \mathcal{X}_1 \rightarrow \mathcal{Y}_1, \dots, Y_n : \mathcal{X}_n \rightarrow \mathcal{Y}_n$ . Then,  $L = L_1 \times \dots \times L_n$ , where  $L_i = \{x : Y_i(x) = y_i\}$  is the section of  $\mathcal{X}_i$  determined by  $y_i$ . Consequently, the complete data and observed data likelihood functions  $f_X(x; \theta)$  and  $f_Y(y; \theta)$  become products, and the corresponding support functions  $\lambda_X(x; \theta)$  and  $\lambda_Y(y; \theta)$  become summations. Substituting these results into (2-3) and interchanging the order of integration and summation gives

$$\theta^{(k+1)} = \arg \max_{\theta} \sum_{i=1}^n E_{\theta^{(k)}}[\lambda_{X_i}(X_i; \theta) | X_i \in L_i] \quad (2-6)$$

for the update of the parameter vector  $\theta^{(k)}$ , where

$$E_{\theta^{(k)}}[\lambda_{X_i}(X_i; \theta) | X_i \in L_i] = \int_{L_i} \lambda_{X_i}(x_i; \theta) f_{X_i|Y_i}(x_i|y_i; \theta^{(k)}) dx_i \quad (2-7)$$

is the conditional expectation of the support function for the complete data vector  $x_i$  at the  $k$ th iteration, and

$$f_{X_i|Y_i}(x_i|y_i; \theta) = \frac{f_{X_i}(x_i; \theta)}{f_{Y_i}(y_i; \theta)} = \frac{f_{X_i}(x_i; \theta)}{\int_{L_i} f_{X_i}(x_i; \theta) dx_i}. \quad (2-8)$$

is the conditional density function of  $x_i$  given the observed data vector  $y_i$ .

## 2.3 MAXIMUM A POSTERIORI ESTIMATION

The EM method can also be used to find the maximum *a posteriori* estimate of  $\theta$  in a Bayesian model for the parameter vector. Let  $\Theta$  denote the random variable associated with  $\theta$ , let  $f_{\Theta}(\theta)$  denote its prior density function, and let  $\lambda_{\Theta}(\theta) = \log f_{\Theta}(\theta)$  denote the prior support function. The posterior observed data support function  $\lambda_{\Theta|Y}(\theta|y) = \log f_{\Theta|Y}(\theta|y)$  is obtained using Bayes' formula:

$$\lambda_{\Theta|Y}(\theta|y) = \lambda_{Y|\Theta}(y|\theta) + \lambda_{\Theta}(\theta) - \lambda_Y(y). \quad (2-9)$$

The observed data support function  $\lambda_Y(y; \theta)$  is written in this expression as  $\lambda_{Y|\Theta}(y|\theta)$  to emphasize Bayesian conditioning rather than parametric dependence on  $\theta$ . The maximum *a posteriori* estimate of  $\theta$ , denoted  $\hat{\theta}$ , is that realization of the random variable  $\Theta$  that maximizes  $f_{\Theta|Y}(\theta|y)$  or, equivalently,  $\lambda_{\Theta|Y}(\theta|y)$ , given the sample  $y$ ; in other words,

$$\hat{\theta} = \arg \max_{\theta} \lambda_{\Theta|Y}(\theta|y) = \arg \max_{\theta} \{ \lambda_{Y|\Theta}(y|\theta) + \lambda_{\Theta}(\theta) \}. \quad (2-10)$$

As discussed in [1], the maximum *a posteriori* estimate of  $\theta$  is obtained using the EM method by solving the following sequence of complete data problems:

$$\theta^{(k+1)} = \arg \max_{\theta} \{ E_{\theta^{(k)}} [ \lambda_{X|\Theta}(X|\theta) | X \in L ] + \lambda_{\Theta}(\theta) \} \quad (2-11)$$

for  $k = 0, 1, \dots$ , where  $\theta^{(0)}$  is an initial estimate of  $\theta$ , and  $\lambda_{X|\Theta}(X|\theta)$  is the complete data support function conditioned on  $\theta$ . The arguments in [1] and [28] imply that each EM iteration increases the value of  $\lambda_{\Theta|Y}(\theta|y)$ . Also, as stated in [1], when  $f_{\Theta}(\theta)$  is a natural conjugate prior density function for  $\Theta$ , the function to be maximized in (2-11) often has the same form as that in (2-3) and, so, can be maximized in the same way. This is indeed the case for the Gaussian mixture models discussed later. (Recall that a natural conjugate prior density function for  $\Theta$  has the property that the posterior density function is a member of the same family of density function. See Redner et al. [31] for a discussion of a natural family of priors, which they refer to as a family of “class conditional” priors, for mixtures of density functions of the exponential type.)



### 3. OBSERVED INFORMATION FOR INCOMPLETE DATA PROBLEMS

Louis's approach to computing the observed information matrix for the general case of incomplete data is presented in this section, as well as a simplification of his result for the special case of independent observations. The posterior observed information matrix, used here as an information measure for stochastic dynamic mixture models, is also defined. Throughout this section, the observation  $y$  is assumed given.

#### 3.1 GENERAL CASE

Before deriving the observed information matrix for the general case, some additional notation and useful identities are presented. (The notation adopted here follows Louis's, with a few differences. His derivation is found in the appendix of [3].) Let  $S_X(x; \theta)$  and  $B_X(x; \theta)$  denote the first derivatives and negative second derivatives with respect to the parameter vector  $\theta$  of the complete data support function  $\lambda_X(x; \theta)$ , respectively. Likewise, let  $S_Y(y; \theta)$  and  $B_Y(y; \theta)$  denote the corresponding derivatives of the observed data support function. (The functions  $S_X(x; \theta)$  and  $S_Y(y; \theta)$  are often referred to as the complete data and observed data score functions, respectively.) These definitions lead to the following identities:

$$S_X(x; \theta) = \lambda'_X(x; \theta) = \frac{f'_X(x; \theta)}{f_X(x; \theta)}, \quad (3-1)$$

$$-B_X(x; \theta) = \lambda''_X(x; \theta) = \frac{f''_X(x; \theta)}{f_X(x; \theta)} - S_X(x; \theta)S_X^T(x; \theta). \quad (3-2)$$

Taking the conditional expectation of these expressions as in (2-4) yields the identities

$$E_\theta \left[ \frac{f'_X(X; \theta)}{f_X(X; \theta)} \mid X \in L \right] = E_\theta [S_X(X; \theta) \mid X \in L], \quad (3-3)$$

$$E_\theta \left[ \frac{f''_X(X; \theta)}{f_X(X; \theta)} \mid X \in L \right] = -E_\theta [B_X(X; \theta) \mid X \in L] + E_\theta [S_X(X; \theta)S_X^T(X; \theta) \mid X \in L]. \quad (3-4)$$

The *information matrix*, denoted  $I_Y(y; \theta)$ , is by definition equal to minus the second derivative with respect to the parameter vector  $\theta$  of the observed data support function:

$$I_Y(y; \theta) = -\lambda''_Y(y; \theta) = B_Y(y; \theta). \quad (3-5)$$

Evaluated at the maximum likelihood estimate  $\hat{\theta}$ , the information matrix is called the *observed information matrix* or, sometimes, the *observed Fisher information matrix*. The *expected (Fisher) information matrix* is defined as the expected value of the information matrix  $I_Y(y; \theta)$  evaluated at the "true" value of the parameter vector  $\theta$ , denoted  $\theta^*$ ; that is,

$$I(\theta^*) = \int_Y I_Y(y; \theta^*) f_Y(y; \theta^*) dy. \quad (3-6)$$

(See Edwards [32] for statements of these definitions.) In practice, the inverse of the expected value of the information matrix evaluated at  $\hat{\theta}$ , that is,  $I^{-1}(\hat{\theta})$ , is often used as an estimate of the error-covariance matrix for  $\hat{\theta}$ . However, Efron and Hinkley [4] give several justifications for preferring the inverse of the observed information matrix, namely,  $I_Y^{-1}(y; \hat{\theta})$ , as a measure of estimation uncertainty, at least for the scalar parameter case. As noted by Louis, the observed information matrix is often much easier to compute than the expected information matrix. This is certainly the case for mixture distributions.

The information matrix  $I_Y(y; \theta)$  as given by (3-5) is often difficult to compute explicitly due to the complexity of the observed data support function  $\lambda_Y(y; \theta)$ . Indeed, it is for this reason that the EM method is often used in the first place. The goal of the EM method is to obtain an expression for the maximum likelihood estimate  $\hat{\theta}$  in terms of the simpler complete data support function  $\lambda_X(x; \theta)$ . Likewise, the goal here is to obtain an expression for  $I_Y(y; \theta)$  in terms of the complete data support function and its derivatives.

To derive the information matrix  $I_Y(y; \theta)$  in terms of the complete data statistics  $S_X(x; \theta)$  and  $B_X(x; \theta)$  requires two steps. The first step is to compute the implicit derivative of the observed data support function. From (2-1),

$$S_Y(y; \theta) = \lambda'_Y(y; \theta) = \int_L f'_X(x; \theta) dx \Big/ \int_L f_X(x; \theta) dx. \quad (3-7)$$

Moving the denominator inside the integral in the numerator and multiplying the numerator and denominator of the integrand by  $f_X(x; \theta)$ , it follows from (2-5) and (3-3) that

$$S_Y(y; \theta) = E_\theta [S_X(X; \theta) | X \in L], \quad (3-8)$$

where the conditional expectation is defined as in (2-4). The second step is to implicitly compute the negative second derivative of the observed data support function. From (3-7),

$$B_Y(y; \theta) = -\lambda''_Y(y; \theta) = - \int_L f''_X(x; \theta) dx \Big/ \int_L f_X(x; \theta) dx + S_Y(y; \theta) S_Y^\top(y; \theta). \quad (3-9)$$

Again, moving the denominator inside the integral in the numerator and multiplying the numerator and denominator of the integrand by  $f_X(x; \theta)$ , it follows from (3-4) and (3-5) that

$$I_Y(y; \theta) = E_\theta [B_X(X; \theta) | X \in L] - E_\theta [S_X(X; \theta) S_X^\top(X; \theta) | X \in L] + S_Y(y; \theta) S_Y^\top(y; \theta). \quad (3-10)$$

By definition, the derivative of the observed data support function is zero at the maximum likelihood estimate, that is,  $S_Y(y; \hat{\theta}) = 0$ . Thus, the observed information matrix in the general case can be expressed entirely in terms of conditional expectations of the complete data statistics  $S_X(x; \theta)$  and  $B_X(x; \theta)$ :

$$I_Y(y; \hat{\theta}) = E_{\hat{\theta}} [B_X(X; \hat{\theta}) | X \in L] - E_{\hat{\theta}} [S_X(X; \hat{\theta}) S_X^\top(X; \hat{\theta}) | X \in R]. \quad (3-11)$$

These expectations are straightforward to compute for the finite and dynamic mixture models considered in this report, and they result in intuitive and revealing expressions for the observed information and error-covariance matrices for these models.

The first term on the right-hand side of (3-11) represents the information in the complete data; the second term is a correction for the information lost to the missing data. To see this, recall expression (2-5) for the conditional density  $f_{X|Y}(x|y; \theta)$  of the complete data random variable  $X$  given  $y$ . Let  $\lambda_{X|Y}(x|y; \theta) = \log f_{X|Y}(x|y; \theta)$ . Taking the natural logarithm of (2-5) and rearranging terms gives

$$\lambda_Y(y; \theta) = \lambda_X(x; \theta) - \lambda_{X|Y}(x|y; \theta). \quad (3-12)$$

Furthermore, taking the conditional expectation as in (2-4) of minus the second derivative of this expression and evaluating the result at the estimate  $\hat{\theta}$  gives

$$I_Y(y; \hat{\theta}) = E_{\hat{\theta}} \left[ B_X(X; \hat{\theta}) \mid X \in L \right] - E_{\hat{\theta}} \left[ -\lambda_{X|Y}(X|y; \hat{\theta}) \mid X \in L \right] \quad (3-13)$$

for the observed information matrix. This result is written succinctly as

$$I_Y(y; \hat{\theta}) = I_X(x; \hat{\theta}) - I_{X|Y}(x|y; \hat{\theta}), \quad (3-14)$$

where, by analogy with definition (3-5), the matrix  $I_X(x; \hat{\theta})$  is called the (conditional expected) complete data observed information matrix, and  $I_{X|Y}(x|y; \hat{\theta})$ , the observed information matrix associated with the missing data. (For brevity, but with some abuse of terminology, these matrices will be referred to as the complete and missing information matrices, respectively.) Clearly, the observed information decreases as the information lost to the missing data increases. Consequently, the error-covariance matrix for the maximum likelihood estimate  $\hat{\theta}$  (taken here to be the inverse of the observed information matrix) increases with the information lost to the missing data. As pointed out by Louis, the factorization (3-14) is an application of what Orchard and Woodbury [33] call the “missing information principle” to the observed information matrix.

### 3.2 INDEPENDENT OBSERVATIONS

In this case of independent observations, the complete data and observed data support functions  $\lambda_X$ ,  $\lambda_Y$  and their first and second derivatives  $S_X$ ,  $S_Y$  and  $-B_X$ ,  $-B_Y$  become summations, and expression (3-11) for the observed information matrix becomes

$$\begin{aligned} I_Y(y; \hat{\theta}) = & \sum_{i=1}^n E_{\hat{\theta}} \left[ B_{X_i}(X_i; \hat{\theta}) \mid X_i \in R_i \right] - \sum_{i=1}^n E_{\hat{\theta}} \left[ S_{X_i}(X_i; \hat{\theta}) S_{X_i}^T(X_i; \hat{\theta}) \mid X_i \in L_i \right] \\ & - 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n E_{\hat{\theta}} \left[ S_{X_i}(X_i; \hat{\theta}) \mid X_i \in L_i \right] E_{\hat{\theta}} \left[ S_{X_j}^T(X_j; \hat{\theta}) \mid X_j \in R_j \right]. \end{aligned} \quad (3-15)$$

This expression from Louis [3] simplifies further in the following way. Since  $S_Y(y; \hat{\theta}) = 0$ , it follows from (3-8) that

$$\sum_{i=1}^n E_{\hat{\theta}} [S_{X_i}(X_i; \hat{\theta}) | X_i \in L_i] = 0. \quad (3-16)$$

Hence, solving for the  $i$ th term,

$$E_{\hat{\theta}} [S_{X_i}(X_i; \hat{\theta}) | X_i \in L_i] = - \sum_{j=1, j \neq i}^n E_{\hat{\theta}} [S_{X_j}(X_j; \hat{\theta}) | X_j \in L_j]. \quad (3-17)$$

By straightforward algebraic manipulation, it is easy to show using these identities that

$$\begin{aligned} E_{\hat{\theta}} [S_X(X; \hat{\theta}) S_X^T(X; \hat{\theta}) | X \in R] &= \sum_{i=1}^n E_{\hat{\theta}} [S_{X_i}(X_i; \hat{\theta}) S_{X_i}^T(X_i; \hat{\theta}) | X_i \in L_i] \\ &\quad - \sum_{i=1}^n E_{\hat{\theta}} [S_{X_i}(X_i; \hat{\theta}) | X_i \in L_i] E_{\hat{\theta}} [S_{X_i}^T(X_i; \hat{\theta}) | X_i \in R_j]. \end{aligned} \quad (3-18)$$

Thus, the observed information matrix in the case of independent observations becomes

$$\begin{aligned} I_Y(y; \hat{\theta}) &= \sum_{i=1}^n E_{\hat{\theta}} [B_{X_i}(X_i; \hat{\theta}) | X_i \in R_i] - \sum_{i=1}^n E_{\hat{\theta}} [S_{X_i}(X_i; \hat{\theta}) S_{X_i}^T(X_i; \hat{\theta}) | X_i \in L_i] \\ &\quad + \sum_{i=1}^n E_{\hat{\theta}} [S_{X_i}(X_i; \hat{\theta}) | X_i \in L_i] E_{\hat{\theta}} [S_{X_i}^T(X_i; \hat{\theta}) | X_i \in R_j]. \end{aligned} \quad (3-19)$$

This simplified expression eliminates the double sum over the cross terms in (3-15).

Finally, for independent and identically distributed data, the observed information matrix  $I_Y(y; \hat{\theta})$  can be approximated by the *empirical Fisher information matrix*, denoted  $I_e(y; \hat{\theta})$ , and defined as  $I_e(y; \hat{\theta}) = n \bar{I}_e(y; \hat{\theta})$ , where

$$\bar{I}_e(y, \theta) = \frac{1}{n} \sum_{i=1}^n S_{Y_i}(y_i; \theta) S_{Y_i}^T(y_i; \theta) - \frac{1}{n^2} S_Y(y; \theta) S_Y^T(y; \theta) \quad (3-20)$$

is the empirical (sample) covariance matrix of the score vectors  $S_{Y_i}(y_i; \theta)$ . (Recall that if the data are identically distributed, the score functions  $S_{Y_i}$  are all the same function.) Since  $S_Y(y; \hat{\theta}) = 0$ , it follows that

$$I_e(y; \hat{\theta}) = \sum_{i=1}^n S_{Y_i}(y_i; \hat{\theta}) S_{Y_i}^T(y_i; \hat{\theta}). \quad (3-21)$$

Using relationship (3-8), the empirical Fisher information matrix can be written in terms of conditional expectations of the complete data score functions:

$$I_e(y; \hat{\theta}) = \sum_{i=1}^n E_{\hat{\theta}} [S_{X_i}(X_i; \hat{\theta}) | X_i \in L_i] E_{\hat{\theta}} [S_{X_i}^T(X_i; \hat{\theta}) | X_i \in L_i]. \quad (3-22)$$



The appropriateness of this approximation to the observed information matrix depends on the size of the data set, as discussed in appendix A. See Redner and Walker [34] and Meilijson [35] for a complete discussion of the empirical Fisher information matrix and its use in the EM method.

### 3.3 POSTERIOR OBSERVED INFORMATION

By analogy with the definition of the information matrix  $I_Y(y; \theta)$ , the *posterior information matrix*  $I_{\Theta|Y}(\theta|y)$  is defined as minus the second derivative with respect to  $\theta$  of the posterior observed data support function  $\lambda_{\Theta|Y}(\theta|y)$ . From (2-9),

$$I_{\Theta|Y}(\theta|y) = -\lambda''_{\Theta|Y}(\theta|y) = -\lambda''_{Y|\Theta}(y|\theta) - \lambda''_{\Theta}(\theta), \quad (3-23)$$

where  $-\lambda''_{Y|\Theta}(y|\theta)$  is the information matrix  $I_Y(y; \theta)$ , denoted  $I_{Y|\Theta}(y|\theta)$  here to emphasize Bayesian conditioning on  $\theta$ , and  $-\lambda''_{\Theta}(\theta)$  is the *prior information matrix*, denoted  $I_{\Theta}(\theta)$ . Hence,

$$I_{\Theta|Y}(\theta|y) = I_{Y|\Theta}(y|\theta) + I_{\Theta}(\theta). \quad (3-24)$$

When evaluated at the maximum *a posteriori* estimate  $\hat{\theta}$ ,  $I_{\Theta|Y}(\hat{\theta}|y)$  and  $I_{\Theta}(\hat{\theta})$  are called the *posterior observed information matrix* and the *prior observed information matrix*, respectively. Consequently, the posterior observed information matrix is equal to the observed information matrix plus the prior observed information matrix. Evaluating (3-24) at  $\hat{\theta}$  and substituting the factorization (3-14) for the observed information matrix gives the following expression for the posterior observed information matrix:

$$I_{\Theta|Y}(\hat{\theta}|y) = I_{X|\Theta}(x|\hat{\theta}) - I_{X|Y,\Theta}(x|y, \hat{\theta}) + I_{\Theta}(\hat{\theta}), \quad (3-25)$$

where the information matrices  $I_{X|\Theta}(x|\theta)$  and  $I_{X|Y,\Theta}(x|y, \theta)$  are the analogs of the complete and missing information matrices  $I_X(x; \theta)$  and  $I_{X|Y}(x|y; \theta)$ , respectively, in the Bayesian model for  $\theta$ . Thus, the posterior observed information matrix can be written entirely in terms of complete data and prior statistics. Again for brevity, but with some abuse of terminology, the combination of the first and last terms in (3-25) will be referred to as the posterior complete information matrix, denoted  $I_{\Theta|X}(\hat{\theta}|x)$ , so that

$$I_{\Theta|Y}(\hat{\theta}|y) = I_{\Theta|X}(\hat{\theta}|x) - I_{X|Y,\Theta}(x|y, \hat{\theta}); \quad (3-26)$$

that is, the posterior observed information matrix is equal to the posterior complete information matrix minus the information lost to the missing data.



## 4. OBSERVED INFORMATION FOR FINITE MIXTURE MODELS

Maximum likelihood estimation and computation of the observed information matrix for finite mixture models (and Gaussian mixture models, in particular) are reviewed in this section. Gaussian mixture models are the basis for the dynamic mixture models considered in later sections.

### 4.1 GENERAL CASE

Let  $y = \{y_i : i = 1, \dots, n\}$  be a given set of  $n$  independent and identically distributed  $p$ -variate observations, assumed to come from a mixture of  $m$  distinct sources in unknown proportions. The goal is to find the maximum likelihood estimates of the parameters for each source distribution in the mixture, and the proportion each source contributes to the data. (In general, the number of sources  $m$  must also be inferred from the data, but that problem is not addressed here.) Finding these estimates would be straightforward if the data were labeled, that is, if each observation  $y_i$  came with a label  $z_i$  taking a value in the set  $\{1, \dots, m\}$  indicating the source from which it came. The labels  $z = \{z_i : i = 1, \dots, n\}$  are the missing data in this problem. The complete data are then  $x = \{x_i : i = 1, \dots, n\}$ , where  $x_i = (y_i, z_i)$  is the complete data vector associated with the observed data vector  $y_i$ .

It is assumed that the labels  $z$  are independent; since the observed data  $y$  are assumed to be independent, it follows that the complete data  $x$  are independent as well. Moreover, since the section  $L_i$  of the complete data sample space  $\mathcal{X}_i$  determined by the observation  $y_i$  is simply the set  $\{y_i\} \times \{1, \dots, m\}$ , it follows that the integrals over  $L_i$  described in the preceding sections are simply sums over the  $m$  possible sources of  $y_i$ . In particular, using the identity

$$f_{X_i}(x_i; \theta) = f_{Y_i Z_i}(y_i, z_i; \theta) = f_{Y_i|Z_i}(y_i|z_i; \theta) f_{Z_i}(z_i; \theta), \quad (4-1)$$

the observed data likelihood function for the sample  $y_i$  is given by the marginal

$$f_{Y_i}(y_i; \theta) = \int_{L_i} f_{X_i}(x_i; \theta) dx_i = \sum_{j=1}^m f_{Y_i|Z_i}(y_i|j; \theta) f_{Z_i}(j; \theta). \quad (4-2)$$

Thus,  $f_{Y_i}(y_i; \theta)$  is a mixture density function, where  $f_{Y_i|Z_i}(y_i|j; \theta)$  is the density function for the sample  $y_i$  given that it comes from source  $j$ , and  $f_{Z_i}(j; \theta)$  is the *a priori* probability of drawing a sample from this source.

Given an estimate  $\theta^{(k)}$  for the mixture parameters  $\theta$ , the updated estimate  $\theta^{(k+1)}$  is obtained by evaluating the conditional expectations (2-7), and maximizing the sum of the results, as in (2-6). Combining the previous two results with the identity

$$f_{X_i|Y_i}(x_i|y_i; \theta) = \frac{f_{X_i}(x_i; \theta)}{f_{Y_i}(y_i; \theta)} = \frac{f_{Y_i Z_i}(y_i, z_i; \theta)}{f_{Y_i}(y_i; \theta)} = f_{Z_i|Y_i}(z_i|y_i; \theta), \quad (4-3)$$

the conditional expectations (2-7) for the finite mixture model become

$$E_{\theta^{(k)}} [\lambda_{X_i}(X_i; \theta) | X_i \in L_i] = \sum_{j=1}^m [\lambda_{Y_i|Z_i}(y_i|j; \theta) + \lambda_{Z_i}(j; \theta)] f_{Z_i|Y_i}(j|y_i; \theta^{(k)}), \quad (4-4)$$

where

$$f_{Z_i|Y_i}(j|y_i; \theta) = \frac{f_{Y_i|Z_i}(y_i|j; \theta) f_{Z_i}(j; \theta)}{\sum_{l=1}^m f_{Y_i|Z_i}(y_i|l; \theta) f_{Z_i}(l; \theta)} \quad (4-5)$$

is the conditional probability that the sample  $y_i$  comes from source  $j$ .

The observed information matrix for the finite mixture model is given by (3-19). The conditional expectations in (3-19) are computed as in (4-4). In particular,

$$E_{\hat{\theta}} [B_{X_i}(X_i; \theta) | X_i \in R_i] = - \sum_{j=1}^m [\lambda_{Y_i|Z_i}(y_i|j; \theta) + \lambda_{Z_i}(j; \theta)]'' f_{Z_i|Y_i}(j|y_i; \hat{\theta}), \quad (4-6)$$

$$E_{\hat{\theta}} [S_{X_i}(X_i; \theta) | X_i \in R_i] = \sum_{j=1}^m [\lambda_{Y_i|Z_i}(y_i|j; \theta) + \lambda_{Z_i}(j; \theta)]' f_{Z_i|Y_i}(j|y_i; \hat{\theta}), \quad (4-7)$$

and

$$E_{\hat{\theta}} [S_{X_i}(X_i; \theta) S_{X_i}^T(X_i; \theta) | X_i \in L_i] = \sum_{j=1}^m [\lambda_{Y_i|Z_i}(y_i|j; \theta) + \lambda_{Z_i}(j; \theta)]' [\lambda_{Y_i|Z_i}(y_i|j; \theta) + \lambda_{Z_i}(j; \theta)]'^T f_{Z_i|Y_i}(j|y_i; \hat{\theta}). \quad (4-8)$$

The expectations (4-6) and (4-7) simplify further by interchanging the order of summation and differentiation:

$$E_{\hat{\theta}} [B_{X_i}(X_i; \theta) | X_i \in R_i] = - \{E_{\hat{\theta}} [\lambda_{X_i}(X_i; \theta) | X_i \in L_i]\}'' , \quad (4-9)$$

$$E_{\hat{\theta}} [S_{X_i}(X_i; \theta) | X_i \in R_i] = \{E_{\hat{\theta}} [\lambda_{X_i}(X_i; \theta) | X_i \in L_i]\}' . \quad (4-10)$$

Therefore, once the expectation (4-4) required to compute  $\hat{\theta}$  is obtained, it need only be differentiated twice to obtain two out of the three expectations required to compute the observed information matrix  $I_Y(y; \hat{\theta})$ , as given by (3-19).

## 4.2 GAUSSIAN MIXTURES

For finite Gaussian (or normal) mixture models, the conditional observed data density function  $f_{Y_i|Z_i}(y_i|z_i; \theta)$  is taken to be the multivariate Gaussian density function  $\phi(y_i|\mu_{z_i}, \Sigma_{z_i})$ , where

$$\phi(a|b, C) = \frac{1}{(2\pi)^{p/2} |C|^{1/2}} \exp \left\{ -\frac{1}{2} (a - b)^T C^{-1} (a - b) \right\} \quad (4-11)$$

is the  $p$ -variate Gaussian density function with mean vector  $b$  and covariance matrix  $C$ . Additionally, the prior probability  $f_{Z_i}(z_i; \theta)$  is taken to be the probability  $\pi_{z_i}$ , where  $\{\pi_1, \dots, \pi_m\}$  is a fixed set of *a priori* probabilities.

### 4.2.1 Parameter Estimation

The parameters to be estimated in this model are, in general, the mean vectors  $\mu_j$ , the covariance matrices  $\Sigma_j$ , and the prior probabilities  $\pi_j$  for the missing measurement-to-source labels. However, to simplify the analysis of the observed information matrix, unequal but known values of the covariance matrices  $\Sigma_j$  are assumed. Hence, the parameter vector  $\theta$  for this problem contains the mean vectors  $\mu_j$  and the mixing proportions  $\pi_j$ . Consequently, the complete data support function  $\lambda_{X_i}(x_i; \theta)$  for this model is

$$\lambda_{X_i}((y_i, z_i); \theta) = -\frac{1}{2}(y_i - \mu_{z_i})^\top \Sigma_{z_i}^{-1}(y_i - \mu_{z_i}) + \log \pi_{z_i}, \quad (4-12)$$

where the first and second terms correspond to the conditional observed data support function  $\lambda_{Y_i|Z_i}(y_i|z_i; \theta)$  and the prior support function  $\lambda_{Z_i}(z_i; \theta)$ , respectively, and terms not dependent on the parameter vector  $\theta$  are dropped from this expression.

It is important to emphasize that the mixing proportions  $\pi_j$  are not independent. In particular,

$$\sum_{j=1}^m \pi_j = \sum_{j=1}^m f_{Z_i}(j; \theta) = 1, \quad \pi_j \geq 0, \quad j = 1, \dots, m, \quad (4-13)$$

and one must be careful to account for this dependence when estimating the mixing proportions and computing the observed information matrix  $I_Y(y; \hat{\theta})$ . In the sequel,  $\pi_m$  is used to denote  $1 - \pi_1 - \dots - \pi_{m-1}$ , and the full expression is employed when taking derivatives of the complete data support function to ensure proper accounting of the constraint in (4-13).

The update equations for the mixing proportions  $\pi_j$  and mean vectors  $\mu_j$  are obtained by substituting the Gaussian mixture model described above into expressions (4-4) and (4-5), and performing the maximization in (2-6) subject to the constraint in (4-13). To simplify notation, let  $w_{ji}$  denote the conditional probability  $f_{Z_i|Y_i}(j|y_i; \theta)$  that observation  $y_i$  comes from source  $j$ . Then, given estimates for  $\pi_j$  and  $\mu_j$  from the  $k$ th iteration, the update equations are

$$\pi_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^n w_{ji}^{(k)}, \quad (4-14)$$

and

$$\mu_j^{(k+1)} = \frac{1}{n\pi_j^{(k+1)}} \sum_{i=1}^n w_{ji}^{(k)} y_i, \quad (4-15)$$

with

$$w_{ji}^{(k)} = \frac{\pi_j^{(k)} \phi(y_i | \mu_j^{(k)}, \Sigma_j)}{\sum_{l=1}^m \pi_l^{(k)} \phi(y_i | \mu_l^{(k)}, \Sigma_l)}. \quad (4-16)$$

### 4.2.2 Observed Information Matrix Computation

The observed information matrix for this case is an  $((m-1) + pm) \times ((m-1) + pm)$  block matrix, where the  $(m-1) \times (m-1)$  block in the upper left-hand corner contains the information contribution from the first  $m-1$  mixing proportions, and the  $pm \times pm$  block in the lower right-hand corner contains the information contribution due to the  $m$  mean vectors, each of length  $p$ . The off-diagonal blocks pertain to information in the various mixing proportion and mean vector combinations. The observed information matrix computation requires computation of the expectations (4-8) through (4-10). Let  $\alpha_j, \beta_l$  denote any two parameters in the set  $\{\pi_1, \dots, \pi_{m-1}, \mu_1, \dots, \mu_m\}$ . To simplify notation, the following shorthand is introduced for expressions (4-8) through (4-10):

$$\langle S_i S_i^\top \rangle_{\alpha_j \beta_l} = E_{\hat{\theta}} [\{\nabla_{\alpha_j} \lambda_{X_i}(X_i; \theta)\} \{\nabla_{\beta_l} \lambda_{X_i}(X_i; \theta)\}^\top | X_i \in L_i], \quad (4-17)$$

$$\langle B_i \rangle_{\alpha_j \beta_l} = -\nabla_{\alpha_j} \{\nabla_{\beta_l} E_{\hat{\theta}} [\lambda_{X_i}(X_i; \theta) | X_i \in L_i]\}^\top, \quad (4-18)$$

$$\langle S_i \rangle_{\alpha_j} = \nabla_{\alpha_j} E_{\hat{\theta}} [\lambda_{X_i}(X_i; \theta) | X_i \in L_i]. \quad (4-19)$$

Substituting the complete data support function (4-12) for the Gaussian mixture into the above expressions yields the following results:

a. From (4-19),

$$\langle S_i \rangle_{\pi_j} = w_{ji}/\pi_j - w_{mi}/\pi_m, \quad j = 1, \dots, m-1, \quad (4-20)$$

$$\langle S_i \rangle_{\mu_j} = w_{ji} \Sigma_j^{-1} (y_i - \mu_j), \quad j = 1, \dots, m. \quad (4-21)$$

b. From (4-18),

$$\langle B_i \rangle_{\pi_j \pi_l} = \begin{cases} w_{ji}/\pi_j^2 + w_{mi}/\pi_m^2, & j = l, \\ w_{mi}/\pi_m^2, & j \neq l, \end{cases} \quad j, l = 1, \dots, m-1, \quad (4-22)$$

$$\langle B_i \rangle_{\mu_j \mu_l} = \begin{cases} w_{ji} \Sigma_j^{-1}, & j = l, \\ 0, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (4-23)$$

$$\langle B_i \rangle_{\pi_j \mu_l} = 0, \quad j = 1, \dots, m-1, \quad l = 1, \dots, m. \quad (4-24)$$

c. From (4-17),

$$\langle S_i S_i^\top \rangle_{\pi_j \pi_l} = \begin{cases} w_{ji}/\pi_j^2 + w_{mi}/\pi_m^2, & j = l, \\ w_{mi}/\pi_m^2, & j \neq l, \end{cases} \quad j, l = 1, \dots, m-1, \quad (4-25)$$

$$\langle S_i S_i^T \rangle_{\mu_j \mu_l} = \begin{cases} w_{ji} \Sigma_j^{-1} (y_i - \mu_j) (y_i - \mu_j)^T \Sigma_j^{-1}, & j = l, \\ 0, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (4-26)$$

$$\langle S_i S_i^T \rangle_{\pi_j \mu_l} = \begin{cases} \frac{w_{ji}}{\pi_j} (y_i - \mu_j)^T \Sigma_j^{-1}, & j, l = 1, \dots, m-1, j = l, \\ 0, & j, l = 1, \dots, m-1, j \neq l, \\ -\frac{w_{mi}}{\pi_m} (y_i - \mu_m)^T \Sigma_m^{-1}, & j = 1, \dots, m-1, l = m. \end{cases} \quad (4-27)$$

The expectations required in (4-25) through (4-27) are obtained using the following results for the first derivatives of the complete data support function:

$$\nabla_{\pi_j} \lambda_{X_i}((y_i, z_i); \theta) = \begin{cases} \frac{1}{\pi_j} & \text{if } z_i = j, \\ -\frac{1}{\pi_m} & \text{if } z_i = m, \\ 0 & \text{otherwise,} \end{cases} \quad (4-28)$$

$$\nabla_{\mu_j} \lambda_{X_i}((y_i, z_i); \theta) = \begin{cases} \Sigma_j^{-1} (y_i - \mu_j) & \text{if } z_i = j, \\ 0 & \text{otherwise.} \end{cases} \quad (4-29)$$

Finally, let  $I_{\hat{\alpha}_j \hat{\beta}_l}$  denote the sub-block of the observed information matrix associated with the parameter estimates  $\hat{\alpha}_j, \hat{\beta}_l$ . Then, from (3-19), using the above shorthand,

$$I_{\hat{\alpha}_j \hat{\beta}_l} = \sum_{i=1}^n \langle B_i \rangle_{\hat{\alpha}_j \hat{\beta}_l} - \sum_{i=1}^n \langle S_i S_i^T \rangle_{\hat{\alpha}_j \hat{\beta}_l} + \sum_{i=1}^n \langle S_i \rangle_{\hat{\alpha}_j} \langle S_i^T \rangle_{\hat{\beta}_l}. \quad (4-30)$$

Substituting (4-20) through (4-27) into (4-30), it follows that the terms in (4-22) cancel with the terms in (4-25). Also, the sum of the terms in (4-27) equals zero when evaluated at the estimates  $\hat{\pi}_j$  and  $\hat{\mu}_j$  as given by (4-14) and (4-15). These results lead to the following simplifications of the observed information matrix for Gaussian mixtures:

$$I_{\hat{\pi}_j \hat{\pi}_l} = \sum_{i=1}^n \langle S_i \rangle_{\hat{\pi}_j} \langle S_i^T \rangle_{\hat{\pi}_l}, \quad j, l = 1, \dots, m-1, \quad (4-31)$$

$$I_{\hat{\pi}_j \hat{\mu}_l} = \sum_{i=1}^n \langle S_i \rangle_{\hat{\pi}_j} \langle S_i^T \rangle_{\hat{\mu}_l}, \quad j = 1, \dots, m-1, \quad l = 1, \dots, m, \quad (4-32)$$

$$I_{\hat{\mu}_j \hat{\mu}_l} = \sum_{i=1}^n \langle S_i \rangle_{\hat{\mu}_j} \langle S_i^T \rangle_{\hat{\mu}_l}, \quad j, l = 1, \dots, m, \quad j \neq l. \quad (4-33)$$

Thus, use of the empirical Fisher information matrix as an approximation to the observed information matrix is somewhat justified in this case, although the extra calculations in (4-23) and (4-26) required to obtain the exact observed information matrix are hardly prohibitive.





## 5. OBSERVED INFORMATION FOR DYNAMIC MIXTURE MODELS

As discussed in the introduction, dynamic mixtures are useful models for data collected over time originating from a number of distinct moving sources. The goal is to estimate the source trajectories, that is, to “track” the sources in the parameter space and to accurately characterize the uncertainty in the track estimates. Intuitively, the uncertainty in the estimated tracks should increase when the sources interfere with each other in the observation space, for example, when two or more trajectories cross paths; measurement-to-source assignment uncertainty is often a significant source of uncertainty in these situations.

In the following sections two important dynamic mixture models are presented, one in which the sources follow unknown but deterministic trajectories, and one in which the trajectories themselves are subject to random perturbations. Gaussian mixtures are used to model the distribution of the observations at each sampling time in both cases. It is assumed that at each time  $t = 1, \dots, T$ , a set of  $n_t$  independent samples  $y_t = \{y_{ti}\}$ ,  $i = 1, \dots, n_t$ , is obtained from the mixture. Let  $y = \{y_t\}$  denote the entire collection of samples, and let  $x_t = \{x_{ti}\}$  and  $x = \{x_t\}$  denote the corresponding sets of complete data samples. It is also assumed that the sets  $x_t$  and  $y_t$  are independent across the sampling times. Note, however, that since the sources in the mixture are in motion, these sets are not identically distributed from one time to the next.

### 5.1 DETERMINISTIC MOTION

In the deterministic case the motion model for each source is embedded in the observation matrix of the standard multivariate linear model. In particular, assuming that the  $p \times 1$  measurement vector  $y_{ti}$  comes from source  $j$ ,  $y_{ti}$  is related to the  $q \times 1$  vector of kinematic parameters  $\mu_j$  through the equation

$$y_{ti} = M_{jt}\mu_j + \epsilon_{jti}, \quad (5-1)$$

where  $M_{jt}$  is a  $p \times q$  matrix that maps  $\mu_j$  to the observation space at time  $t$ , and  $\epsilon_{jti}$  is a  $p \times 1$  Gaussian random vector with zero mean and covariance matrix  $R_{jt}$ . The random errors  $\epsilon_{jti}$  are assumed independent.

For example, suppose  $m$  sources move with constant velocity in a plane, and observations of the source positions are obtained at multiple sampling times. The trajectory of source  $j$  is completely specified by its position and velocity at an arbitrary reference time  $t_*$ . Let  $\mu_j$  be the  $xy$ -position and  $xy$ -velocity of source  $j$  at time  $t_*$ . The position of source  $j$  at any time  $t$  is given by  $M_{jt}\mu_j$ , where

$$M_{jt} = \begin{bmatrix} 1 & 0 & \Delta_{t,t_*} & 0 \\ 0 & 1 & 0 & \Delta_{t,t_*} \end{bmatrix} \quad (5-2)$$

and  $\Delta_{t,t_*}$  is the elapsed time between  $t$  and  $t_*$ .

Given the linear Gaussian model (5-1) for observation  $y_{ti}$  assuming source  $j$ , the conditional observed data density function  $f_{Y_{ti}|Z_{ti}}(y_{ti}|z_{ti}; \theta)$  is the multivariate Gaussian density function  $\phi(y_{ti}|M_{jt}\mu_j, R_{jt})$ . As before, the prior probability  $f_{Z_{ti}}(z_{ti}; \theta)$  of observing source  $j$  at time  $t$  is taken to be the probability  $\pi_{z_i}$ , where  $\{\pi_1, \dots, \pi_m\}$  is a fixed set of *a priori* probabilities for observing data from each source.

### 5.1.1 Parameter Estimation

The parameters to be estimated in this model are, in general, the kinematic parameter vectors  $\mu_j$  and the measurement covariance matrices  $R_j$ , together with the prior probabilities  $\pi_j$  for the missing measurement-to-source assignments. However, to simplify the analysis of the observed information matrix for this case, unequal but known values of the covariance matrices  $R_j$  are assumed. Thus, the parameter vector  $\theta$  contains the kinematic vectors  $\mu_j$  and the mixing proportions  $\pi_j$ . Consequently, the complete data support function  $\lambda_{X_i}(x_i; \theta)$  for this model is

$$\lambda_{X_{ti}}((y_{ti}, z_{ti}); \theta) = \left[ -\frac{1}{2}(y_{ti} - M_{jt}\mu_j)^T R_j^{-1}(y_{ti} - M_{jt}\mu_j) + \log \pi_j \right] \Big|_{j=z_{ti}}, \quad (5-3)$$

where the first and second terms in brackets correspond to the conditional support function  $\lambda_{Y_{ti}|Z_{ti}}(y_{ti}|z_{ti}; \theta)$  and the prior support function  $\lambda_{Z_{ti}}(z_{ti}; \theta)$ , respectively, and terms not dependent on  $\theta$  are dropped from this expression for clarity.

The update equations for the mixing proportions  $\pi$  and the kinematic parameters  $\mu$  are obtained by substituting the linear Gaussian dynamic mixture model described above into the analogs of expressions (4-4) and (4-5) for data collected over more than one sampling time, and maximizing the resulting expressions with respect to  $\pi$  and  $\mu$  as in (2-6) subject to the constraint (4-13); the single sum over the measurement index  $i$  in (2-6) is replaced by the double sum over the time and measurement indices  $t$  and  $i$ , respectively, in this case. Let  $w_{jti}$  denote the conditional probability  $f_{Z_{ti}|Y_{ti}}(j|y_{ti}; \theta)$  that observation  $y_{ti}$  comes from source  $j$ . Then, given estimates for  $\pi_j$  and  $\mu_j$  from the  $k$ th EM iteration, the update equations for the conditional probabilities and mixing proportions are

$$w_{jti}^{(k)} = \frac{\pi_j^{(k)} \phi(y_{ti}|M_{jt}\mu_j^{(k)}, R_{jt})}{\sum_{l=1}^m \pi_l^{(k)} \phi(y_{ti}|M_{lt}\mu_l^{(k)}, R_{lt})}, \quad (5-4)$$

and

$$\pi_j^{(k+1)} = \frac{1}{n} \sum_{t=1}^T \sum_{i=1}^{n_t} w_{jti}^{(k)}, \quad (5-5)$$

where  $n = \sum_{t=1}^T n_t$  is the total number of observations over all sampling times.

The update equations for the kinematic parameters  $\mu_j$  assume intuitively appealing forms when written in terms of the “synthetic” measurements\*

$$\tilde{y}_{jt}^{(k)} = \frac{\sum_{i=1}^{n_t} w_{jti}^{(k)} y_{ti}}{\sum_{i=1}^{n_t} w_{jti}^{(k)}}, \quad (5-6)$$

and synthetic measurement covariance matrices

$$\tilde{R}_{jt}^{(k)} = \frac{R_{jt}}{\sum_{i=1}^{n_t} w_{jti}^{(k)}}. \quad (5-7)$$

The synthetic measurement  $\tilde{y}_{jt}$  for source  $j$  at time  $t$  is the probabilistic centroid of the observations  $y_{ti}$  with respect to the assignment probabilities  $w_{jti}$ . The synthetic measurement covariance matrix  $\tilde{R}_{jt}$  is the measurement covariance matrix for source  $j$  at time  $t$  divided by the expected number of measurements from this source, conditioned on the observations  $y_{ti}$ . To see this, define the indicator functions

$$1_j(x_{ti}) = 1_j((y_{ti}, z_{ti})) = \begin{cases} 1, & \text{if } z_{ti} = j, \\ 0, & \text{otherwise,} \end{cases} \quad (5-8)$$

and let  $n_{jt}(x_t) = \sum_{i=1}^{n_t} 1_j(x_{ti})$  be the number of measurements that come from source  $j$  at time  $t$ . Then,

$$E[n_{jt}(X_t) | X_t \in L_t] = \sum_{l=1}^m \sum_{i=1}^{n_t} 1_j((y_{ti}, l)) w_{l ti} = \sum_{i=1}^{n_t} w_{j ti}. \quad (5-9)$$

is the expected number of observations  $y_{ti}$  that come from source  $j$ . Incidentally, the *a priori* expected number of observations from source  $j$  at time  $t$  is

$$E[n_{jt}(X_t)] = \sum_{l=1}^m \sum_{i=1}^{n_t} 1_j((y_{ti}, l)) \pi_l = n_t \pi_j. \quad (5-10)$$

Using expressions (5-6) and (5-7) for the synthetic measurements and synthetic measurement covariance matrices for source  $j$ , the update equation for the kinematic parameter vector  $\mu_j$  is

$$\left( \sum_{t=1}^T M_{jt}^T [\tilde{R}_{jt}^{(k)}]^{-1} M_{jt} \right) \mu_j^{(k+1)} = \left( \sum_{t=1}^T M_{jt}^T [\tilde{R}_{jt}^{(k)}]^{-1} \tilde{y}_{jt}^{(k)} \right). \quad (5-11)$$

This set of linear equations to be solved for  $\mu_j$  is the set of normal equations for  $\mu_j$  from linear least-squares theory. It follows that the updated estimate for  $\mu_j$  is the weighted least-squares estimate for  $\mu_j$  given the synthetic measurements  $\tilde{y}_{jt}$  with weights determined by the synthetic measurement covariance matrices  $\tilde{R}_{jt}$ .

---

\*The term “synthetic” in this context is adopted from Streit and Luginbuhl [8].

### 5.1.2 Observed Information Matrix Computation

The observed information matrix for the linear Gaussian dynamic mixture model is similar to that for the standard Gaussian mixture model as given by expressions (4-20) through (4-33), but with the subscripts for the measurement index  $i$  replaced with subscripts for the time and measurement indices  $t$  and  $i$ , respectively, and the single sums over  $i$  replaced with double sums over  $t$  and  $i$ . In particular, the observed information matrix for this case is an  $((m-1) + qm) \times ((m-1) + qm)$  block matrix, where the  $(m-1) \times (m-1)$  block in the upper left-hand corner contains the information contribution from the first  $m-1$  mixing proportions, and the  $qm \times qm$  block in the lower right-hand corner contains the information contribution due to the  $m$  kinematic parameter vectors, each of length  $q$ . Substituting the complete data support function (5-3) into the analogs of expressions (4-17) through (4-19) for data collected over multiple sampling times gives the following results:

a. From the time-dependent form of (4-19),

$$\langle S_{ti} \rangle_{\pi_j} = w_{jti}/\pi_j - w_{mti}/\pi_m, \quad j = 1, \dots, m-1, \quad (5-12)$$

$$\langle S_{ti} \rangle_{\mu_j} = w_{jti} M_{jt}^T R_{jt}^{-1} (y_{ti} - M_{jt} \mu_j), \quad j = 1, \dots, m. \quad (5-13)$$

b. From the time-dependent form of (4-18),

$$\langle B_{ti} \rangle_{\pi_j \pi_l} = \begin{cases} w_{jti}/\pi_j^2 + w_{mti}/\pi_m^2, & j = l, \\ w_{mti}/\pi_m^2, & j \neq l, \end{cases} \quad j, l = 1 \dots, m-1, \quad (5-14)$$

$$\langle B_{ti} \rangle_{\mu_j \mu_l} = \begin{cases} w_{jti} M_{jt}^T R_{jt}^{-1} M_{jt}, & j = l, \\ 0, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (5-15)$$

$$\langle B_{ti} \rangle_{\pi_j \mu_l} = 0, \quad j = 1 \dots, m-1, \quad l = 1, \dots, m. \quad (5-16)$$

c. From the time-dependent form of (4-17),

$$\langle S_{ti} S_{ti}^T \rangle_{\pi_j \pi_l} = \begin{cases} w_{jti}/\pi_j^2 + w_{mti}/\pi_m^2, & j = l, \\ w_{mti}/\pi_m^2, & j \neq l, \end{cases} \quad j, l = 1, \dots, m-1, \quad (5-17)$$

$$\langle S_{ti} S_{ti}^T \rangle_{\mu_j \mu_l} = \begin{cases} w_{jti} M_{jt}^T R_{jt}^{-1} (y_{ti} - M_{jt} \mu_j) (y_{ti} - M_{jt} \mu_j)^T R_{jt}^{-1} M_{jt}, & j = l, \\ 0, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (5-18)$$

$$\langle S_{ti} S_{ti}^\top \rangle_{\pi_j \mu_l} = \begin{cases} \frac{w_{jti}}{\pi_j} (y_{ti} - M_{jt} \mu_j)^\top R_{jt}^{-1} M_{jt}, & j, l = 1, \dots, m-1, j = l, \\ 0, & j, l = 1, \dots, m-1, j \neq l, \\ -\frac{w_{mti}}{\pi_m} (y_{ti} - M_{mt} \mu_m)^\top R_{mt}^{-1} M_{mt}, & j = 1, \dots, m-1, l = m. \end{cases} \quad (5-19)$$

As before, let  $\alpha_j, \beta_l$  denote any two parameters in the set  $\{\pi_1, \dots, \pi_{m-1}, \mu_1, \dots, \mu_m\}$ , and let  $I_{\hat{\alpha}_j \hat{\beta}_l}$  denote the sub-block of the observed information matrix associated with the parameter estimates  $\hat{\alpha}_j, \hat{\beta}_l$ . Then, from the time-dependent form of (3-19), using the above shorthand,

$$I_{\hat{\alpha}_j \hat{\beta}_l} = \sum_{t=1}^T \sum_{i=1}^{n_t} \langle B_{ti} \rangle_{\hat{\alpha}_j \hat{\beta}_l} - \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} S_{ti}^\top \rangle_{\hat{\alpha}_j \hat{\beta}_l} + \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} \rangle_{\hat{\alpha}_j} \langle S_{ti}^\top \rangle_{\hat{\beta}_l}. \quad (5-20)$$

Substituting (5-12) through (5-19) into (5-20), it follows that the terms in (5-14) cancel with the terms in (5-17), and that the sum of the terms in (5-19) equals zero when evaluated at the estimates  $\hat{\pi}_j$  and  $\hat{\mu}_j$ , as given by (5-5) and (5-11). These results lead to the following simplifications of the observed information matrix for linear Gaussian dynamic mixtures:

$$I_{\hat{\pi}_j \hat{\pi}_l} = \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} \rangle_{\hat{\pi}_j} \langle S_{ti}^\top \rangle_{\hat{\pi}_l}, \quad j, l = 1, \dots, m-1, \quad (5-21)$$

$$I_{\hat{\pi}_j \hat{\mu}_l} = \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} \rangle_{\hat{\pi}_j} \langle S_{ti}^\top \rangle_{\hat{\mu}_l}, \quad j = 1, \dots, m-1, \quad l = 1, \dots, m, \quad (5-22)$$

$$I_{\hat{\mu}_j \hat{\mu}_l} = \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} \rangle_{\hat{\mu}_j} \langle S_{ti}^\top \rangle_{\hat{\mu}_l}, \quad j, l = 1, \dots, m, \quad j \neq l. \quad (5-23)$$

Note, however, that use of the empirical Fisher information matrix as an approximation to the observed information matrix is not strictly speaking justified in this case, as the observations  $y_{ti}$  are not identically distributed across sampling times. Indeed, the mixture distribution for the observations  $y_{ti}$  in general changes location and shape from one time to the next due to the motion of the sources.

### 5.1.3 Interpretation of Complete Information Matrix

Before proceeding to stochastic motion models for dynamic mixtures, it is worth examining the statistical interpretation of the complete information matrix for the deterministic case. The first term in (5-20) is the sub-block of the complete information matrix associated with the estimates  $\hat{\alpha}_j, \hat{\beta}_l$ ; the last two terms represent the information lost to the missing data. Let  $[I_X]_{\hat{\mu} \hat{\mu}}$  denote the  $qm \times qm$  block of the complete information matrix associated with the collection of kinematic vectors  $\mu_j$ , let  $[I_X]_{\hat{\mu}_j \hat{\mu}_j}$  denote the  $j$ th diagonal  $q \times q$  sub-block of this matrix, and let  $\{u_j : j = 1, \dots, m\}$  be the collection of unit vectors of length  $m$ , where the

$j$ th element of  $u_j$  equals one and all other elements of  $u_j$  equal zero. Substituting (5-15) into the first term of (5-20) and using the synthetic measurement covariance matrices (5-7) gives

$$[I_X]_{\hat{\mu}\hat{\mu}} = \sum_{j=1}^m u_j u_j^T \otimes [I_X]_{\hat{\mu}_j \hat{\mu}_j} = \sum_{j=1}^m u_j u_j^T \otimes \sum_{t=1}^T M_{jt}^T \tilde{R}_{jt}^{-1} M_{jt}. \quad (5-24)$$

This result may be interpreted in terms of the M-step at the final EM iteration (that is, iteration  $k = \infty$ ) as follows. For each  $j \in \{1, \dots, m\}$ , consider the multivariate linear model

$$\tilde{y}_{jt} = M_{jt} \mu_j + \gamma_{jt}, \quad t = 1, \dots, T, \quad (5-25)$$

where  $\tilde{y}_{jt}$  is a  $p \times 1$  measurement vector,  $M_{jt}$  is a known  $p \times q$  observation matrix,  $\mu_j$  is a  $q \times 1$  parameter vector to be estimated, and  $\gamma_{jt}$  are  $p \times 1$  independent, normally distributed noise vectors with zero means and known covariance matrices  $\tilde{R}_{jt}$ . Then,

$$\hat{\mu}_{(MVU)j} = \left( \sum_{t=1}^T M_{jt}^T \tilde{R}_{jt}^{-1} M_{jt} \right)^{-1} \left( \sum_{t=1}^T M_{jt}^T \tilde{R}_{jt}^{-1} \tilde{y}_{jt} \right) \quad (5-26)$$

is the minimum variance unbiased (MVU) estimate for  $\mu_j$  with error-covariance matrix

$$C_{\hat{\mu}_{(MVU)j}} = \left( \sum_{t=1}^T M_{jt}^T \tilde{R}_{jt}^{-1} M_{jt} \right)^{-1}, \quad (5-27)$$

assuming that this matrix has full rank. The inverse of this matrix is the Fisher information matrix for this problem. Comparing (5-26) with (5-11), it follows that the M-step at the final EM iteration is equivalent to the MVU estimation problem for the multivariate linear model (5-25) with independent measurements  $\tilde{y}_{jt}^{(\infty)}$  and known measurement covariance matrices  $\tilde{R}_{jt}^{(\infty)}$ . Furthermore, from (5-27) and (5-24), it follows that the complete information matrix for  $\hat{\mu}_j$  is equivalent to the Fisher information matrix for  $\mu_j$  for this MVU estimation problem; that is,

$$[I_X]_{\hat{\mu}_j \hat{\mu}_j} = C_{\hat{\mu}_{(MVU)j}}^{-1}. \quad (5-28)$$

Thus, the observed information matrix for  $\hat{\mu}_j$  can be written using the missing information principle as in (3-14) as follows:

$$[I_Y]_{\hat{\mu}_j \hat{\mu}_j} = C_{\hat{\mu}_{(MVU)j}}^{-1} - [I_{X|Y}]_{\hat{\mu}_j \hat{\mu}_j}, \quad (5-29)$$

where  $[I_{X|Y}]_{\hat{\mu}_j \hat{\mu}_j}$  is the information lost to the missing data. This seemingly obvious connection between the complete information matrix obtained at the final EM iteration and the Fisher information matrix obtained from the equivalent MVU estimation problem for this deterministic dynamic mixture model leads to a clearer understanding of the analogous result for the stochastic model discussed in the next section.

## 5.2 STOCHASTIC MOTION

Suppose the trajectory of each source is subject to random perturbations about an underlying deterministic motion model. Then, each source trajectory may be treated as a sequence of random variables for which the deterministic motion model is the mean. In particular, let  $\mu_{jt}$  denote the kinematic “state” of source  $j$  at time  $t$ . Then, it is assumed that the states  $\mu_j = \{\mu_{jt} : t = 0, 1, \dots, T\}$  are related by the first-order Gauss-Markov process

$$\mu_{jt} = F_{jt,t-1}\mu_{j,t-1} + \delta_{j,t-1}, \quad (5-30)$$

where  $F_{jt,t-1}$  are known  $q \times q$  state transition matrices, and  $\delta_{jt}$  are independent  $q \times 1$  Gaussian random vectors with zero means and known covariance matrices  $Q_{jt}$ . Additionally, the state of source  $j$  at time  $t_0$  is assumed to be normally distributed with mean  $\eta_j$  and covariance matrix  $\Gamma_j$ . As for the deterministic case, it is assumed that the observations  $y_{ti}$  are related linearly to the source states  $\mu_{jt}$  so that, assuming that observation  $y_{ti}$  comes from source  $j$ ,

$$y_{ti} = M_{jt}\mu_{jt} + \epsilon_{jti}, \quad (5-31)$$

where again  $M_{jt}$  are known  $p \times q$  observation matrices, and  $\epsilon_{jti}$  are independent  $p \times 1$  Gaussian random vectors with zero means and known covariance matrices  $R_{jt}$ . Furthermore, it is assumed that the random perturbations  $\delta_{jt}$  and  $\epsilon_{jti}$  are independent.

Consider again the constant-velocity motion example presented for the deterministic case. As before, suppose that  $m$  sources move (nominally) with constant velocity in a plane, and that observations of source positions are obtained at multiple sampling times, but that the kinematic state vectors for the sources are subject to random perturbations between sampling times. Then, to within these perturbations, the position and velocity of source  $j$  at time  $t$  can be predicted based on the position and velocity of the source at time  $t - 1$  using the state transition matrix  $F_{jt,t-1}$ . For the nominal constant-velocity model, the predicted  $xy$ -position and  $xy$ -velocity of source  $j$  at time  $t$  is  $F_{jt,t-1}\mu_{j,t-1}$ , with state transition matrix

$$F_{jt,t-1} = \begin{bmatrix} 1 & 0 & \Delta_{t,t-1} & 0 \\ 0 & 1 & 0 & \Delta_{t,t-1} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (5-32)$$

The position of source  $j$  at time  $t$  is simply  $M_{jt}\mu_{jt}$ , with observation matrix

$$M_{jt} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \quad (5-33)$$

With the distributional assumptions on the source states  $\mu = \{\mu_j : j = 1, \dots, m\}$  in the Gauss-Markov process model (5-30), and assuming a diffuse prior for the mixing proportions  $\pi = \{\pi_j : j = 1, \dots, m\}$ , the joint density function for the parameters  $\theta = (\pi, \mu)$  is

$$f_{\Theta}((\pi, \mu)) \propto \prod_{j=1}^m \phi(\mu_{j0} | \eta_j, \Gamma_j) \prod_{t=1}^T \phi(\mu_{jt} | F_{jt,t-1} \mu_{j,t-1}, Q_{j,t-1}), \quad (5-34)$$

where the sources are assumed to be independent, and the states for each source are conditionally independent from one sampling time to the next. The complete data and observed data density functions for this stochastic dynamic mixture model are essentially the same as for the deterministic model, except for the dependence of the state vectors  $\mu_{jt}$  on the sampling times. Specifically, the complete data density function is

$$f_{X|\Theta}((y, z) | \theta) = \prod_{t=1}^T \prod_{i=1}^{n_t} [\pi_j \phi(y_{ti} | M_{jt} \mu_{jt}, R_{jt})] |_{j=z_{ti}}. \quad (5-35)$$

Marginalizing over the missing measurement-to-source assignments  $z$ , and interchanging the order of the sums and products, gives the observed data density function

$$f_{Y|\Theta}(y | \theta) = \prod_{t=1}^T \prod_{i=1}^{n_t} \sum_{j=1}^m \pi_j \phi(y_{ti} | M_{jt} \mu_{jt}, R_{jt}). \quad (5-36)$$

### 5.2.1 State Estimation

The parameters to be estimated in this model are the mixing proportions  $\pi$  and the kinematic state vectors  $\mu$ . The state vectors  $\mu_j = \{\mu_{jt} : t = 0, 1, \dots, T\}$  for source  $j$  are treated as a concatenated state (column) vector for the rest of this discussion. The system matrices  $\{F_{jt,t-1}\}$ ,  $\{Q_{jt}\}$ ,  $\{M_{jt}\}$ , and  $\{R_{jt}\}$  are all assumed to be known.

The update equations for the maximum *a posteriori* estimates of  $\pi$  and  $\mu$  are obtained by substituting (5-34) and (5-35) into (2-11) and performing the necessary expectations and maximizations. Let

$$\Psi(\theta | \theta^{(k)}) = E_{\theta^{(k)}}[\lambda_{X|\Theta}(X | \theta) | X \in L] + \lambda_{\Theta}(\theta) \quad (5-37)$$

denote the E-step at the  $k$ th EM iteration. Substituting (5-34) and (5-35) into this expression and performing the expectation gives

$$\begin{aligned} \Psi(\theta | \theta^{(k)}) = & \sum_{j=1}^m \left\{ \log \phi(\mu_{j0} | \eta_j, \Gamma_j) + \sum_{t=1}^T \log \phi(\mu_{jt} | F_{jt,t-1} \mu_{j,t-1}, Q_{j,t-1}) \right. \\ & \left. + \sum_{t=1}^T \sum_{i=1}^{n_t} w_{ji}^{(k)} \log \phi(y_{ti} | M_{jt} \mu_{jt}, R_{jt}) \right\} + \sum_{j=1}^m \sum_{t=1}^T \sum_{i=1}^{n_t} w_{ji}^{(k)} \log \pi_j, \quad (5-38) \end{aligned}$$



where the conditional probabilities  $w_{jti} = f_{Z_{ti}|Y_{ti},\Theta}(z_{ti}|y_{ti},\theta)$  are obtained from the ratio of (5-35) and (5-36):

$$w_{jti}^{(k)} = \frac{\pi_j^{(k)} \phi(y_{ti}|M_{jt}\mu_{jt}^{(k)}, R_{jt})}{\sum_{l=1}^m \pi_l^{(k)} \phi(y_{ti}|M_{lt}\mu_{lt}^{(k)}, R_{lt})}. \quad (5-39)$$

Since the prior distribution for the mixing proportions  $\pi_j$  is assumed to be uninformative (that is, uniform), the update equations for  $\pi_j$  are identical to the update equations (5-5) for the deterministic case.

The update equations for the kinematic state vectors  $\mu_j$  are more complicated for the stochastic case because of the time dependence between states due to the Markov model (5-30). As for the deterministic case, these update equations assume intuitively appealing forms when the E-step is written in terms of the synthetic measurements (5-6) and synthetic measurement covariance matrices (5-7). After some tedious algebraic manipulation, and ignoring an additive constant that does not depend on  $\pi$  or  $\mu$ , the result is

$$\begin{aligned} \Psi(\theta|\theta^{(k)}) = & \sum_{j=1}^m \left\{ \log \phi(\mu_{j0}|\eta_j, \Gamma_j) + \sum_{t=1}^T \log \phi(\mu_{jt}|F_{jt,t-1}\mu_{j,t-1}, Q_{j,t-1}) \right. \\ & \left. + \sum_{t=1}^T \log \phi(\tilde{y}_{jt}^{(k)}|M_{jt}\mu_{jt}, \tilde{R}_{jt}^{(k)}) \right\} + \sum_{j=1}^m \sum_{t=1}^T \sum_{i=1}^{n_t} w_{jti}^{(k)} \log \pi_j. \end{aligned} \quad (5-40)$$

Let  $\{e_t^\circ : t = 0, 1, \dots, T\}$  be the collection of unit vectors of length  $T + 1$ , where the  $t$ th element of  $e_t^\circ$  equals one and all other elements of  $e_t^\circ$  equal zero, and let  $E_{t\tau}^\circ = e_t^\circ e_\tau^{\circ\top}$  for all  $t, \tau = 0, 1, \dots, T$ . Taking the derivative of (5-40) with respect to the  $\mu_j$  and setting the result equal to zero gives the following  $q(T + 1) \times q(T + 1)$  system of equations for source  $j$ :

$$\left[ I_{(data)j}^{(k)} + I_{(prior)j} \right] \mu_j^{(k+1)} = \left[ d_{(data)j}^{(k)} + d_{(prior)j} \right], \quad (5-41)$$

where

$$I_{(data)j}^{(k)} = \sum_{t=1}^T E_{tt}^\circ \otimes M_{jt}^\top [\tilde{R}_{jt}^{(k)}]^{-1} M_{jt}, \quad (5-42)$$

$$d_{(data)j}^{(k)} = \sum_{t=1}^T e_t^\circ \otimes M_{jt}^\top [\tilde{R}_{jt}^{(k)}]^{-1} \tilde{y}_{jt}^{(k)}, \quad (5-43)$$

and

$$\begin{aligned} I_{(prior)j} = & E_{00}^\circ \otimes \Gamma_j^{-1} + \sum_{t=1}^T E_{tt}^\circ \otimes Q_{j,t-1}^{-1} + \sum_{t=0}^{T-1} E_{tt}^\circ \otimes F_{j,t+1,t}^\top Q_{jt}^{-1} F_{j,t+1,t} \\ & - \sum_{t=1}^T E_{t-1,t}^\circ \otimes F_{jt,t-1}^\top Q_{j,t-1}^{-1} - \sum_{t=1}^T E_{t,t-1}^\circ \otimes Q_{j,t-1}^{-1} F_{jt,t-1}, \end{aligned} \quad (5-44)$$

$$d_{(prior)j} = e_0^\circ \otimes \Gamma_j^{-1} \eta_j. \quad (5-45)$$

The linear system of equations (5-41) can be efficiently solved for  $\mu_j^{(k+1)}$  using a specialized form of Gaussian elimination for block-tridiagonal systems. Alternatively, as shown by Streit and Luginbuhl in [8], the system can be solved efficiently using a fixed interval Kalman smoothing filter. To see this, observe that the term in braces in expression (5-40) is the natural logarithm of the joint density function for the random state sequence  $\mu_j$  with Gauss-Markov process model (5-30), and observations  $\tilde{y}_{jt}$  with measurement model

$$\tilde{y}_{jt} = M_{jt}\mu_{jt} + \gamma_{jt}, \quad t = 1, \dots, T, \quad (5-46)$$

where  $\gamma_{jt}$  are independent  $p \times 1$  normally distributed noise vectors with zero means and known covariance matrices  $\tilde{R}_{jt}$ . The joint density function for the combined model (5-30) and (5-46) is the joint density function for the fixed interval smoothing problem of Kalman filtering theory. (A useful reference on Kalman filtering theory for this work is the book by Mendel [36].) Hence, the state estimates  $\hat{\mu}_j$  obtained from the M-step at the final EM iteration ( $k = \infty$ ) are equivalent to the minimum mean-squared error (MMSE) estimates for  $\mu_j$  obtained from the fixed interval Kalman smoothing filter given the synthetic measurements  $\tilde{y}_{jt}^{(\infty)}$  and synthetic measurement covariance matrices  $\tilde{R}_{jt}^{(\infty)}$ .

The linear Gauss-Markov dynamic mixture model presented in this section is precisely the tracking model used in the PMHT method of Streit and Luginbuhl [8]. In their report, the authors attempt to interpret the Fisher information matrix for the states  $\mu_j$  in terms of the error-covariance matrices obtained at the output of the equivalent Kalman smoothing filter for the M-step at the final EM iteration. Their interpretation is not theoretically, by their own admission, completely satisfactory. At the end of this section, an exact statistical interpretation of these matrices is given in terms of the complete information matrix for the state estimates  $\hat{\mu}_j$ . Specifically, it is shown that the error-covariance matrices from the Kalman smoothing filter for  $\hat{\mu}_j$  in the PMHT model are the diagonal blocks of the inverse of the posterior complete information matrix  $I_{\Theta|X}(\hat{\theta}|x)$  corresponding to  $\hat{\mu}_j$ . Consequently, these error-covariance matrices do not account for the information lost to the missing data and, thus, are overly optimistic estimates of estimation error.

### 5.2.2 Posterior Observed Information Matrix Computation

The posterior observed information matrix  $I_{\Theta|Y}(\hat{\theta}|y)$  for the linear Gauss-Markov dynamic mixture is, by (3-24), the sum of the observed information matrix  $I_{Y|\Theta}(y|\hat{\theta})$  for the linear Gaussian mixture measurement model (5-36), and the prior observed information matrix  $I_{\Theta}(\hat{\theta})$  for the Gauss-Markov process model (5-34). The observed information matrix  $I_{Y|\Theta}(y|\hat{\theta})$  is similar to that for the deterministic case, as given by expressions (5-12) through (5-23), except that the sub-block for each state vector  $\mu_j$  is itself a block matrix, with sub-

blocks corresponding to the kinematic state of the source at times  $t = 0, 1, \dots, T$ . In particular, the observed information matrix for this case is an  $((m-1) + q(T+1)m) \times ((m-1) + q(T+1)m)$  block matrix, where the  $(m-1) \times (m-1)$  block in the upper left-hand corner contains the information contribution from the first  $m-1$  mixing proportions, and the  $q(T+1)m \times q(T+1)m$  block in the lower right-hand corner contains the information contribution from the  $m$  concatenated state vectors  $\mu_j$ , each of length  $q(T+1)$ . Substituting the complete data support function obtained from (5-35) into the analogs of expressions (4-17) through (4-19) for data collected over multiple sampling times gives the following computations for the information matrix  $I_{Y|\Theta}(y|\theta)$ :

a. From the time-dependent form of (4-19),

$$\langle S_{ti} \rangle_{\pi_j} = w_{jti}/\pi_j - w_{mti}/\pi_m, \quad j = 1, \dots, m-1, \quad (5-47)$$

$$\langle S_{ti} \rangle_{\mu_j} = e_t^\circ \otimes w_{jti} M_{jt}^\top R_{jt}^{-1} (y_{ti} - M_{jt} \mu_{jt}), \quad j = 1, \dots, m. \quad (5-48)$$

b. From the time-dependent form of (4-18),

$$\langle B_{ti} \rangle_{\pi_j \pi_l} = \begin{cases} w_{jti}/\pi_j^2 + w_{mti}/\pi_m^2, & j = l, \\ w_{mti}/\pi_m^2, & j \neq l, \end{cases} \quad j, l = 1, \dots, m-1, \quad (5-49)$$

$$\langle B_{ti} \rangle_{\mu_j \mu_l} = \begin{cases} E_{tt}^\circ \otimes w_{jti} M_{jt}^\top R_{jt}^{-1} M_{jt}, & j = l, \\ 0, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (5-50)$$

$$\langle B_{ti} \rangle_{\pi_j \mu_l} = 0, \quad j = 1, \dots, m-1, \quad l = 1, \dots, m. \quad (5-51)$$

c. From the time-dependent form of (4-17),

$$\langle S_{ti} S_{ti}^\top \rangle_{\pi_j \pi_l} = \begin{cases} w_{jti}/\pi_j^2 + w_{mti}/\pi_m^2, & j = l, \\ w_{mti}/\pi_m^2, & j \neq l, \end{cases} \quad j, l = 1, \dots, m-1, \quad (5-52)$$

$$\langle S_{ti} S_{ti}^\top \rangle_{\mu_j \mu_l} = \begin{cases} E_{tt}^\circ \otimes w_{jti} M_{jt}^\top R_{jt}^{-1} (y_{ti} - M_{jt} \mu_{jt}) (y_{ti} - M_{jt} \mu_{jt})^\top R_{jt}^{-1} M_{jt}, & j = l, \\ 0, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (5-53)$$

$$\langle S_{ti} S_{ti}^\top \rangle_{\pi_j \mu_l} = \begin{cases} e_t^\circ \otimes \frac{w_{jti}}{\pi_j} (y_{ti} - M_{jt} \mu_{jt})^\top R_{jt}^{-1} M_{jt}, & j, l = 1, \dots, m-1, j = l, \\ 0, & j, l = 1, \dots, m-1, j \neq l, \\ -e_t^\circ \otimes \frac{w_{mti}}{\pi_m} (y_{ti} - M_{mt} \mu_{mt})^\top R_{mt}^{-1} M_{mt}, & j = 1, \dots, m-1, l = m. \end{cases} \quad (5-54)$$

Recall that the prior information matrix  $I_{\Theta}(\theta)$  is the negative second derivative of the prior support function  $\lambda_{\Theta}(\theta) = \log f_{\Theta}(\theta)$ . Hence, from (5-34),

$$-\nabla_{\pi_j} \{\nabla_{\pi_l} \lambda_{\Theta}(\theta)\}^T = 0, \quad j, l = 1, \dots, m-1, \quad (5-55)$$

$$-\nabla_{\mu_j} \{\nabla_{\mu_l} \lambda_{\Theta}(\theta)\}^T = \begin{cases} I_{(prior)j}, & j = l, \\ 0, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (5-56)$$

$$-\nabla_{\pi_j} \{\nabla_{\mu_l} \lambda_{\Theta}(\theta)\}^T = 0, \quad j = 1, \dots, m-1, \quad l = 1, \dots, m. \quad (5-57)$$

Let  $\alpha_j, \beta_l$  denote any two parameters in the set  $\{\pi_1, \dots, \pi_{m-1}, \mu_0, \mu_1, \dots, \mu_m\}$ , and let  $I_{\hat{\alpha}_j \hat{\beta}_l}$  denote the sub-block of the posterior observed information matrix associated with the estimates  $\hat{\alpha}_j, \hat{\beta}_l$ . Then, from (3-24) and the time-dependent form of (3-19), using the above shorthand,

$$I_{\hat{\alpha}_j \hat{\beta}_l} = \sum_{t=1}^T \sum_{i=1}^{n_t} \langle B_{ti} \rangle_{\hat{\alpha}_j \hat{\beta}_l} - \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} S_{ti}^T \rangle_{\hat{\alpha}_j \hat{\beta}_l} + \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} \rangle_{\hat{\alpha}_j} \langle S_{ti}^T \rangle_{\hat{\beta}_l} - \nabla_{\hat{\alpha}_j} \left\{ \nabla_{\hat{\beta}_l} \lambda_{\Theta}(\hat{\theta}) \right\}^T. \quad (5-58)$$

Substituting (5-47) through (5-57) into (5-58), it follows that the terms in (5-49) cancel with the terms in (5-52). These results lead to the following simplifications of the posterior observed information matrix for linear Gauss-Markov dynamic mixtures:

$$I_{\hat{\pi}_j \hat{\pi}_l} = \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} \rangle_{\hat{\pi}_j} \langle S_{ti}^T \rangle_{\hat{\pi}_l}, \quad j, l = 1, \dots, m-1, \quad (5-59)$$

$$I_{\hat{\mu}_j \hat{\mu}_l} = \sum_{t=1}^T \sum_{i=1}^{n_t} \langle S_{ti} \rangle_{\hat{\mu}_j} \langle S_{ti}^T \rangle_{\hat{\mu}_l}, \quad j, l = 1, \dots, m, \quad j \neq l. \quad (5-60)$$

Again, use of the empirical Fisher information matrix as an approximation to the observed information matrix is not appropriate in this case, since the observations  $y_{ti}$  are not identically distributed across sampling times due to source motion.

### 5.2.3 Interpretation of PMHT Error-Covariance Matrices

Finally, the connection between the error-covariance matrices for the state estimates  $\hat{\mu}_j$  obtained from the equivalent Kalman smoothing filters for the M-step at the final EM iteration for the linear Gauss-Markov mixture model and the posterior complete information matrix for these estimates needs to be established. The first and last terms in (5-58) constitute the sub-block of the posterior complete information matrix associated with the estimates  $\hat{\alpha}_j$  and  $\hat{\beta}_l$ ; the middle two terms in this expression represent the information lost to the missing data.

Let  $[I_{\Theta|X}]_{\hat{\mu}\hat{\mu}}$  denote the  $q(T+1)m \times q(T+1)m$  block of the complete information matrix associated with all of the kinematic state vectors, and let  $[I_{\Theta|X}]_{\hat{\mu}_j\hat{\mu}_j}$  denote the  $j$ th diagonal  $q(T+1) \times q(T+1)$  sub-block of this matrix. Substituting (5-50) and (5-56) into the first and last terms of (5-58) and using the synthetic measurement covariance matrices (5-7) gives

$$[I_{\Theta|X}]_{\hat{\mu}\hat{\mu}} = \sum_{j=1}^m u_j u_j^\top \otimes [I_{\Theta|X}]_{\hat{\mu}_j\hat{\mu}_j} = \sum_{j=1}^m u_j u_j^\top \otimes [I_{(data)j} + I_{(prior)j}]. \quad (5-61)$$

This result may be interpreted in terms of the equivalent Kalman smoothing filters for the M-step at the final EM iteration ( $k = \infty$ ) as follows. Consider the concatenated form of the Kalman smoothing model. In particular, let  $\{e_t : t = 1, \dots, T\}$  be the collection of unit vectors of length  $T$ , where the  $t$ th element of  $e_t$  equals one and all other elements of  $e_t$  equal zero, and let  $E_{t\tau} = e_t e_\tau^\top$  for all  $t, \tau = 1, \dots, T$ . Let  $\tilde{y}_j = \sum_{t=1}^T e_t \otimes \tilde{y}_{jt}$  be the concatenated synthetic measurement vector for source  $j$ . Then,

$$\tilde{y}_j = M_j \mu_j + \gamma_j, \quad (5-62)$$

where

$$M_j = \begin{bmatrix} 0 & \sum_{t=1}^T E_{tt} \otimes M_{jt} \end{bmatrix} \quad (5-63)$$

is the corresponding concatenated observation matrix,  $\gamma_j$  is a normally distributed concatenated noise vector with zero mean and known covariance matrix  $\tilde{R}_j = \sum_{t=1}^T E_{tt} \otimes \tilde{R}_{jt}$ , and the concatenated state vector  $\mu_j$  is normally distributed with mean vector

$$v_j = \sum_{t=0}^T e_t^\circ \otimes v_{jt} \quad (5-64)$$

and covariance matrix

$$P_j = \sum_{t=0}^T \sum_{\tau=0}^T E_{t\tau}^\circ \otimes P_{jt\tau}, \quad (5-65)$$

given by the following recursions from Theorem 15-5 in Mendel [36, pp.#217–218]:

$$v_{jt} = \begin{cases} \eta_j, & t = 0, \\ F_{jt,t-1} v_{j,t-1}, & t = 1, \dots, T, \end{cases} \quad (5-66)$$

and

$$P_{jtt} = \begin{cases} \Gamma_j, & t = 0, \\ F_{jt,t-1} P_{j,t-1,t-1} F_{jt,t-1}^\top + Q_{j,t-1}, & t = 1, \dots, T, \end{cases} \quad (5-67)$$

$$P_{jt\tau} = \begin{cases} F_{jt\tau} P_{j\tau\tau}, & t > \tau, \\ P_{jtt} F_{j\tau t}^\top, & t < \tau, \end{cases} \quad t, \tau = 0, 1, \dots, T, \quad t \neq \tau, \quad (5-68)$$

where

$$F_{jt\tau} = F_{jt,t-1}F_{j,t-1,t-2} \cdots F_{j,\tau+1,\tau} \quad \text{for } t > \tau. \quad (5-69)$$

From Theorem 13-2 in Mendel [36, p.#180], the MMSE estimate for  $\mu_j$  for this linear Gaussian model is given by

$$\hat{\mu}_{(MMSE)j} = v_j + P_j M_j^\top (M_j P_j M_j^\top + \tilde{R}_j)^{-1} (\tilde{y}_j - M_j v_j), \quad (5-70)$$

with associated error-covariance matrix

$$P_{\hat{\mu}_{(MMSE)j}} = (P_j^{-1} + M_j^\top \tilde{R}_j^{-1} M_j)^{-1}. \quad (5-71)$$

It is straightforward to show that

$$M_j^\top \tilde{R}_j^{-1} M_j = I_{(data)j}. \quad (5-72)$$

Furthermore, it is shown in appendix B that

$$P_j^{-1} = I_{(prior)j}. \quad (5-73)$$

Thus, from (5-61),

$$P_{\hat{\mu}_{(MMSE)j}} = (P_j^{-1} + M_j^\top \tilde{R}_j^{-1} M_j)^{-1} = [I_{(data)j} + I_{(prior)j}]^{-1} = [I_{\Theta|X}]_{\hat{\mu}_j \hat{\mu}_j}^{-1}; \quad (5-74)$$

that is, the inverse of the posterior complete information matrix for the estimate  $\hat{\mu}_j$  is equal to the error-covariance matrix for the MMSE estimate for  $\mu_j$  given the synthetic measurements and synthetic measurement covariance matrices at the final EM iteration. Moreover, since the fixed interval Kalman smoothing filter is just an efficient algorithm for obtaining the MMSE estimates (5-70) and the diagonal blocks of the error-covariance matrix (5-71), it follows that the error-covariance matrices for the smoothed states obtained from this filter are the diagonal blocks of the inverse of the posterior complete information matrix for the estimate  $\hat{\mu}_j$ .

The posterior observed information matrix for  $\hat{\mu}_j$  can be written using the missing information principle as in (3-26). From (3-26) and (5-74),

$$[I_{\Theta|Y}]_{\hat{\mu}_j \hat{\mu}_j} = P_{\hat{\mu}_{(MMSE)j}}^{-1} - [I_{X|Y,\Theta}]_{\hat{\mu}_j \hat{\mu}_j}, \quad (5-75)$$

where  $[I_{X|Y,\Theta}]_{\hat{\mu}_j \hat{\mu}_j}$  is the information lost to the missing data. Thus, while it is tempting to interpret the error-covariance matrices from the Kalman smoothing filters for the M-step of the final EM iteration as *the* error-covariance matrices for the states estimates  $\hat{\mu}_j$ , it is clear from this expression that these matrices provide only part of the information required to compute error-covariance matrices for  $\hat{\mu}_j$ . In short, the error-covariance matrices obtained

from the Kalman smoothing filters do not account for the information lost to the missing measurement-to-source assignments. It is also clear from (5-75) that the posterior observed information matrix for  $\hat{\mu}_j$  requires computation of the *full* MMSE covariance matrix (5-71), and not just the diagonal blocks provided by the Kalman smoothing filter.

Furthermore, in general, the error-covariance matrix for  $\hat{\mu}_j$  must be taken from the inverse of the *entire* posterior observed information matrix  $I_{\Theta|Y}$  for *all* estimated source states  $\hat{\mu}$  and their mixing proportions  $\hat{\pi}$ , because in general  $I_{\Theta|Y}$  is *not* block-diagonal; that is,

$$[I_{\Theta|Y}^{-1}]_{\hat{\mu}_j \hat{\mu}_j} \neq [I_{\Theta|Y}]_{\hat{\mu}_j \hat{\mu}_j}^{-1}. \quad (5-76)$$

Only in the case of no assignment uncertainty does this inequality become an equality. Indeed, from expressions (5-49) through (5-51), (5-55) through (5-57), and (5-58), it follows that as the information in the missing data (the contribution from the second and third terms in (5-58)) approaches zero, as when the sources move farther apart, the posterior observed information matrix approaches the posterior complete information matrix, which is block-diagonal, so that from (3-26), (5-74), and (5-75),

$$[I_{\Theta|Y}^{-1}]_{\hat{\mu}_j \hat{\mu}_j} \rightarrow [I_{\Theta|X}^{-1}]_{\hat{\mu}_j \hat{\mu}_j} = [I_{\Theta|X}]_{\hat{\mu}_j \hat{\mu}_j}^{-1} = P_{\hat{\mu}(MMSE)_j}. \quad (5-77)$$

Subsequently, when there is no missing data, that is, when the measurement-to-source assignments are known, the diagonal blocks of the inverse of the posterior observed information matrix, or the error-covariance matrices for the state estimates  $\hat{\mu}_j$ , are equal to the error-covariance matrices obtained from the Kalman smoothing filter, which is expected given the assumptions on the distributions of the measurements and the states. Moreover, the inverse of the posterior observed information matrix for the estimates  $\hat{\mu}_j$  is equal to the posterior Cramér-Rao lower bound for the states  $\mu_j$  in this case.





## 6. THEORETICAL AND PRACTICAL CONSIDERATIONS

At least two issues need to be considered when using the inverse of the observed information matrix as an estimate of the error-covariance matrix for dynamic mixture models: the accuracy of the normal approximation to the distribution of  $\hat{\theta}$ , and the cost of computing the inverse. These issues are discussed below.

### 6.1 ASYMPTOTIC NORMALITY OF $\hat{\theta}$

Recall that the asymptotic distribution of the maximum likelihood estimate  $\hat{\theta}$  is normal with mean vector  $\theta^*$  and covariance matrix  $I^{-1}(\theta^*)$ , where  $\theta^*$  is the “true” value of the parameter vector  $\theta$ , and  $I(\theta^*)$  is the Fisher information matrix. There are two obvious estimators for the asymptotic error-covariance matrix  $I^{-1}(\theta^*)$ , namely,  $I^{-1}(\hat{\theta})$  and  $I_Y^{-1}(y; \hat{\theta})$ . In [4], Efron and Hinkley give theory, examples, and evidence from Fisher’s original writings supporting a preference for the estimator  $I_Y^{-1}(y; \hat{\theta})$  over  $I^{-1}(\hat{\theta})$  for scalar parameter families. The simple example at the beginning of their paper succinctly illustrates their reasoning. In any event, both estimators are inferentially valid only for large sample size  $n$ . However, with regard to the scalar parameter examples presented in their paper, Efron and Hinkley note that “repeated sampling, with  $n$  as low as 10, seems to induce normality of the likelihood rather quickly.” On the other hand, McLachlan and Peel [37] state that “the sample size  $n$  has to be very large before the asymptotic theory applies to mixture models.” Determining sufficient sample sizes for appropriate use of these large sample approximations to the error-covariance matrix for the mixture models examined in this report requires further investigation.

In a Bayesian model for  $\theta$ , the distribution of the maximum a posteriori estimate  $\hat{\theta}$  depends on the sample size  $n$  and the nature of the prior distribution. Asymptotically, as  $n \rightarrow \infty$ , the distribution of  $\hat{\theta}$  approaches the distribution of the maximum likelihood estimate for  $\theta$  discussed above. For finite sample sizes, the distribution of  $\hat{\theta}$  depends on the relative strengths of the data and the prior. If the prior is relatively weak, the distribution of  $\hat{\theta}$  will be closer to that of the maximum likelihood estimate. On the other hand, in the absence of data, the distribution of  $\hat{\theta}$  is equivalent to the prior distribution for  $\theta$ .

### 6.2 SEQUENTIAL VERSUS BATCH PROCESSING

Depending on the number of sources  $m$  and sampling times  $T$ , the posterior observed information matrix for stochastic dynamic mixture models can be costly to invert. For example, the number of parameters to be estimated in the linear Gauss-Markov mixture model grows roughly linearly with  $m$  and  $T$ . Specifically, the observed information matrix for this model has dimension  $((m-1) + q(T+1)m) \times ((m-1) + q(T+1)m)$ , where  $q$  is the length of the state vector for each source. Suppose, for instance, the  $xy$ -positions of two sources

( $m = 2$ ) moving with constant velocity ( $q = 4$ ) are observed at 10 different sampling times ( $T = 10$ ). The posterior observed information matrix for the one independent mixing proportion and both sets of state vectors in this case has dimension  $89 \times 89$ . Computing the inverse of this matrix requires roughly  $89^3 \approx 700 \times 10^3$  operations. There are perhaps efficient methods for obtaining this inverse, or selected portions of this inverse (for example, the diagonal elements), but investigations of such methods are beyond the scope of this report.

An alternative but suboptimal approach for computing the posterior observed information matrix for stochastic dynamic mixture models is to reduce the size of the matrix by processing the data sequentially. In this approach, the data collected at each sampling time are processed as if they were the only data collected, and the state estimates and error-covariance matrices computed at this time are used as the mean vectors and covariance matrices of the prior distributions for the states at the next sampling time. The estimates obtained in this way are suboptimal in that they are conditioned only on the data collected up to the current sampling time, and not the entire data set. In the language of Kalman filtering theory, estimates obtained by processing the data sequentially are called filtered estimates; those obtained by conditioning on the entire batch of data are referred to as smoothed estimates. The reduction in the number of computations required to compute error-covariance matrices in this suboptimal filtering approach can be substantial. For the example given above, the posterior observed information matrix for the one independent mixing proportion and the state estimates for the two sources at each sampling time has dimension  $17 \times 17$ . Computing the inverses of these matrices for each of the sampling times requires roughly  $10 \cdot 17^3 \approx 50 \times 10^3$  operations, a reduction by an order of magnitude over the optimal smoothing approach.

While the error-covariance matrices for the state estimates are cheaper to compute using the filtering approach described above, the savings come at the expense of accuracy in both the state estimates and the error-covariance matrices. This is true even for the state estimates and error-covariance matrices obtained at the final sampling time  $T$ , for which one would think smoothing would have no impact. When there is no measurement-to-source assignment uncertainty (for instance, when the sources are widely separated), the filtered estimates and the smoothed estimates of the source states at time  $T$ , and the associated error-covariance matrices, are identical. The EM iterations for this case degenerate to a single iteration that, in terms of the equivalent Kalman smoothers, corresponds to one forward-backward pass over the synthetic measurements for times  $t = 1, \dots, T$ . However, when there is significant interference between the sources, many EM iterations may be required for the state estimates to converge to their final values; each iteration corresponds to a forward-backward pass over the synthetic measurements, whose values change with each pass according to the updated conditional measurement-to-source assignment probabilities.

In practice, a balance between the filtering and smoothing approaches can be achieved by implementing the algorithm as a “sliding” batch. In this approach, the algorithm is first run at each sampling time on a batch of data that is expanded from one sampling time to the next until it reaches a fixed length. When this length is reached, the batch is then slid forward at each new sampling time, so that the data at the current sampling time are added to the batch, and the data from the oldest sampling time in the batch are removed from the batch. Let  $\rho(t)$  denote the batch length at time  $t$ , and let  $\bar{\rho}$  denote the fixed batch length. Then, the batch length at time  $t$  is given by

$$\rho(t) = \begin{cases} t, & t < \bar{\rho}, \\ \bar{\rho}, & \bar{\rho} \leq t \leq T. \end{cases} \quad (6-1)$$

Several authors have proposed similar approaches (see, for example, Rago et al. [38] and Willett et al. [39]), and most have noted that the prior distributions for the states in each batch must be determined in such a way so that they are not functions of data in the current batch. The prior distributions for the states in the sliding batch proposed here are determined as follows. In the expanding stage, the prior mean vectors and covariance matrices specified at time  $t = 0$  are used for each batch. In the sliding stage, the state estimates and error-covariance matrices from the batch at time  $t - \bar{\rho}$  are used in the prior distributions for the batch at time  $t$ ; these error-covariance matrices are computed from the inverse of the posterior observed information matrix for all the sources for the batch at time  $t - \bar{\rho}$ . This approach fixes a problem with PMHT not identified in [39]. In particular, it would appear that the sliding batch approach proposed in [39] uses the error-covariance matrices obtained from the inverse of the posterior complete information matrix as priors for successive batches; it was shown in the previous section that these error-covariance matrices are too small when there is significant measurement-to-source assignment uncertainty.



## 7. EXAMPLES

Two target tracking examples using the linear Gauss-Markov mixture model (that is, the PMHT model) are presented in this section. The first example is of two constant-velocity crossing targets. This example is idealized in the sense that there are no missed detections ( $P_D = 1$ ) and no false alarms ( $P_{FA} = 0$ ). The second example is of a single constant-velocity target in clutter ( $P_{FA} > 0$ ). This example is further complicated by the possibility of missed detections ( $P_D < 1$ ). In each case, the consistency of the target state estimates is examined. As described in the next section, consistency in this context is a measure of how well the estimated error-covariance matrices reflect the actual errors in the state estimates.

### 7.1 ESTIMATOR CONSISTENCY

#### 7.1.1 Parametric Test

Let  $\hat{\mu}_j(t|t)$  denote the state estimate of target  $j$  at time  $t$  given a batch of measurements of length  $\rho(t) \geq 1$  with leading edge at time  $t$  and trailing edge at time  $t - \rho(t)$ , and let  $C_j(t|t)$  denote the corresponding error-covariance matrix. When there is no assignment uncertainty (for instance, when the target measurements are labeled, or when the targets are widely separated) and under the linear Gauss-Markov model, the posterior distribution of the state  $\mu_{jt}$  given the batch of measurements  $y_{t-\rho(t)}, \dots, y_t$  is the normal distribution with mean vector  $\hat{\mu}_j(t|t)$  and covariance matrix  $C_j(t|t)$ . Let  $\tilde{\mu}_j(t|t) = \mu_{jt} - \hat{\mu}_j(t|t)$  denote the estimation error. Under these assumptions, it follows that

$$E[\tilde{\mu}_j(t)|y_{t-\rho(t)}, \dots, y_t] = 0, \quad (7-1)$$

$$\text{cov}(\tilde{\mu}_j(t)|y_{t-\rho(t)}, \dots, y_t) = C_j(t|t). \quad (7-2)$$

A state estimator is said to be *consistent* if the estimation errors have these two properties. Said another way, a state estimator is consistent if the estimation errors have zero mean, and their covariances equal the estimated covariances. (See Bar-Shalom and Li [40] for a full discussion of estimator consistency.)

Recall from [40] that the normalized estimation error squared (NEES) for target  $j$  at time  $t$  is defined as

$$\nu_j(t) = \tilde{\mu}_j^\top(t|t) C_j^{-1}(t|t) \tilde{\mu}_j(t|t). \quad (7-3)$$

Given the modeling assumptions and ideal conditions described above, the NEES  $\nu_j(t)$  is chi-square distributed with mean (degrees of freedom)  $q$ , where  $q$  is the length of the state vector  $\mu_{jt}$ . Suppose the tracking simulation is run  $N$  times. Then, one can compute the average

NEES for each target:

$$\bar{\nu}_j(t) = \frac{1}{N} \sum_{l=1}^N \nu_j^{(l)}(t), \quad t = 1, \dots, T, \quad (7-4)$$

where  $\nu_j^{(l)}(t)$  is the NEES for target  $j$  at time  $t$  for the  $l$ th run. By the properties of chi-square distributed random variables, it follows that  $N$  times the average NEES is chi-square distributed with  $Nq$  degrees of freedom. Hence, to test for estimator consistency, one can test the simple hypothesis

$$H_0 : N\bar{\nu}_j(t) \text{ is chi-square distributed with mean } Nq \quad (7-5)$$

at each time  $t$ . This is a two-sided test, the alternative hypothesis  $H_1$  being that the average NEES  $\bar{\nu}_j(t)$  has mean less than or greater than  $Nq$ . The critical region for this test for a fixed size (level of significance)  $\alpha$  is typically taken to be the lower and upper tails of the chi-square distribution, each with probability mass  $\alpha/2$ . Let  $\chi_\xi^2(r)$  denote the point in the interval  $[0, \infty)$  such that the left-tail probability of the chi-square distribution with degrees of freedom  $\xi$  is  $r$ . Then, the acceptance region (complement of the critical region) for this two-sided test is the interval  $[\chi_{Nq}^2(\alpha/2), \chi_{Nq}^2(1 - \alpha/2)]$ . Simply put, if the null hypothesis  $H_0$  is true, then on average  $(1 - \alpha)\%$  of the average NEES values  $\bar{\nu}_j(t)$ ,  $t = 1, \dots, T$ , will fall within the acceptance region.

### 7.1.2 Nonparametric Test

The test for estimator consistency based on the average NEES value  $\bar{\nu}_j(t)$  is standard in the tracking literature [40]. For comparison, an alternative test based on the sample (or empirical) distribution function of the NEES values  $\nu_j^{(1)}(t), \dots, \nu_j^{(N)}(t)$  is proposed. For detailed discussions of tests of fit based on the empirical distribution function (EDF), see Cramér [29, section 30.8], D'Agostino and Stephens [41], and Stuart et al. [42, sections 25.35–25.44]. For the remainder of this discussion, consider an arbitrary but fixed target  $j$  at an arbitrary but fixed time  $t$ . Let  $\zeta_{(l)}$  denote the  $l$ th order statistic of the NEES values  $\nu_j^{(1)}(t), \dots, \nu_j^{(N)}(t)$ , so that  $\zeta_{(1)} \leq \dots \leq \zeta_{(N)}$ . The EDF for this sample is defined by

$$F_N(\zeta) = \begin{cases} 0, & \zeta < \zeta_{(1)}, \\ l/N, & \zeta_{(l)} \leq \zeta < \zeta_{(l+1)}, \\ 1, & \zeta_{(N)} \leq \zeta. \end{cases} \quad (7-6)$$

There are several statistics based on the EDF used to test against the hypothesized distribution of the sample. The most well known is the Kolmogorov (K) statistic

$$D_N = \sup_{\zeta} |F_N(\zeta) - F(\zeta)|, \quad (7-7)$$

where  $F(\zeta)$  is the true (hypothesized) distribution function for the sample (in this case, the chi-square distribution with mean  $q$ ). Less well known are those from the Cramér-von Mises family of statistics

$$Q_N = \int_{-\infty}^{\infty} (F_N(\zeta) - F(\zeta))^2 \psi(\zeta) dF(\zeta), \quad (7-8)$$

where  $\psi(\zeta)$  are non-negative weighting functions. Of these, the two most studied are the statistic corresponding to  $\psi(\zeta) = 1$ , called the Cramér-von Mises (CVM) statistic, denoted  $W_N$ , and the statistic corresponding to  $\psi(\zeta) = [F(\zeta)(1 - F(\zeta))]^{-1}$ , called the Anderson-Darling (AD) statistic, denoted  $A_N$ . Each of the statistics  $D_N$ ,  $W_N$ , and  $A_N$  is a distance measure between the EDF  $F_N(\zeta)$  and the hypothesized distribution function  $F(\zeta)$ , and each of these statistics can be used in a test-of-fit with simple hypothesis

$$H_0 : \nu_j^{(1)}(t), \dots, \nu_j^{(N)}(t) \text{ come from a chi-square distribution with mean } q. \quad (7-9)$$

This test is usually cast as a one-sided (upper-tail) test. (See [41, section 4.5.1] for the reasoning behind this.) Percentage points (critical values) for these statistics for various significance levels  $\alpha$  are given in [41, table 4.2, p.#150] and in [42, p.#420]; the null hypothesis  $H_0$  is rejected when these values are exceeded.

Properties of these non-parametric statistics are discussed at length in [41]. In summary,  $D_N$  is often much less powerful than  $W_N$  and  $A_N$ , meaning that tests based on  $D_N$  often have a lower probability of accepting the alternative hypothesis  $H_1$  when  $H_1$  is true than tests based on  $W_N$  and  $A_N$ ; each of these statistics is sensitive to deviation from the mean of the hypothesized distribution  $F(\zeta)$ ;  $A_N$  often behaves similarly to  $W_N$ , but is usually more powerful when the EDF  $F_N(\zeta)$  deviates from the hypothesized distribution  $F(\zeta)$  in the tails.

Computation of the EDF statistics  $D_N$ ,  $W_N$ , and  $A_N$  is typically accomplished using the probability integral transformation  $v_j^{(l)}(t) = F(\nu_j^{(l)}(t))$ . In particular, if  $F$  is the true distribution function for the random variables  $\nu_j^{(l)}(t)$ , then the random variables  $v_j^{(l)}(t)$  are uniformly distributed between 0 and 1, and the original test-of-fit becomes a test-of-fit between the EDF for the transformed variables  $v_j^{(l)}(t)$  and the standard uniform distribution function. Let  $v_{(l)}$  denote the  $l$ th order statistic of the values  $v_j^{(1)}(t), \dots, v_j^{(N)}(t)$ , so that  $v_{(1)} \leq \dots \leq v_{(N)}$ . Then, the statistics  $D_N$ ,  $W_N$ , and  $A_N$  are given by

$$D_N = \max \left\{ \max_l \left\{ \frac{l}{N} - v_{(l)} \right\}, \max_l \left\{ v_{(l)} - \frac{l-1}{N} \right\} \right\}, \quad (7-10)$$

$$W_N = \frac{1}{12n} + \sum_{l=1}^N \left[ v_{(l)} - \frac{2l-1}{2N} \right]^2, \quad (7-11)$$

$$A_N = -N - \frac{1}{N} \sum_{l=1}^N [(2l-1) \log v_{(l)} + (2N+1-2l) \log(1-v_{(l)})]. \quad (7-12)$$

For the examples presented here, the values  $v_j^{(l)}(t)$  are obtained by evaluating the chi-square distribution function for  $q$  degrees of freedom at the NEES values  $\nu_j^{(l)}(t)$  for each target  $j$  at each time  $t$  for simulations  $l = 1, \dots, N$ . Proofs related to the probability integral transformation are found in [41, chapter 6] and Stuart and Ord [43, section 1.27].

## 7.2 TWO CROSSING TARGETS

In this example, two constant-velocity targets cross in the  $xy$ -plane; that is, the two targets share the same  $xy$ -position at some time  $t_c$ . (Such a scenario is possible, for instance, when two aircraft cross paths at different altitudes.) At time  $t = 0$ , targets 1 and 2 are at  $xy$ -positions  $(1, 0)$  and  $(2, 0)$ , with  $xy$ -velocities  $(0.05, 2)$  and  $(-0.05, 2)$ , respectively. It follows that the targets cross paths at time  $t_c = 10$ . A single measurement of each target's  $xy$ -position is obtained at each time  $t = 1, \dots, 25$ , for a total of 50 observations over the entire scenario. These measurements have a standard deviation of 0.075 in each dimension. The distance between the two targets in the  $x$ -dimension in units of measurement standard deviation is shown in figure 1. For consistency with the assumption that a single measurement of each target is obtained at each time, the mixing proportions  $\pi_1$  and  $\pi_2$  are each set to 0.5 and held fixed for the simulation. Finally, the mean vector for the prior distribution for each target is taken to be the true position and velocity vector of each target at time  $t = 0$ , so that  $\eta_1 = (1, 0, 0.05, 2)$  and  $\eta_2 = (2, 0, -0.05, 2)$ . The prior covariance and process noise covariance matrices for each target are taken to be

$$\Gamma_j = \text{diag}((1, 1, 0.1, 0.1)) \quad (7-13)$$

and

$$Q_{jt} = 10^{-9} \text{diag}((1, 1, 0.1, 0.1)) \quad (7-14)$$

for  $j = 1, 2$ , and  $t = 0, \dots, 24$ , respectively, where  $\text{diag}(v)$  is the diagonal matrix with the elements of the vector  $v$  on the diagonal.

This simulation was run 100 times for each of four batch lengths:  $\bar{\rho} = 25, 10, 5$ , and 1. For each run, the NEES values (7-3) were computed twice, once using the posterior observed information matrix, that is, with

$$C_j^{-1}(t|t) = [I_{\Theta|Y}]_{\hat{\mu}_j(t|t)}, \quad (7-15)$$

and once using the posterior complete information matrix, that is, with

$$C_j^{-1}(t|t) = [I_{\Theta|X}]_{\hat{\mu}_j(t|t)}, \quad (7-16)$$

where  $[I_{\Theta|Y}]_{\hat{\mu}_j(t|t)}$  and  $[I_{\Theta|X}]_{\hat{\mu}_j(t|t)}$  denote the sub-blocks of the posterior observed and posterior complete information matrices, respectively, associated with the state estimate  $\hat{\mu}_j(t|t)$ .

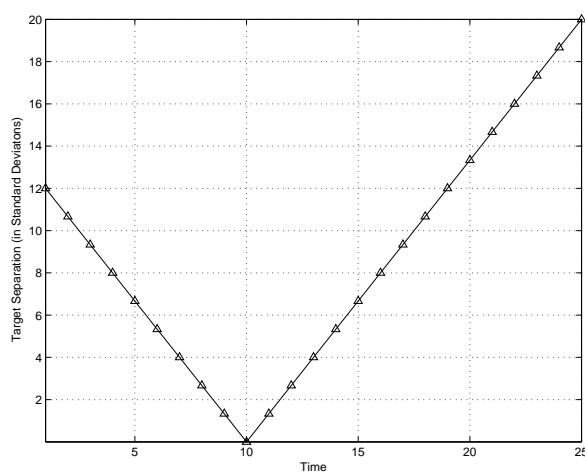


In each case, the average NEES values (7-4) over the 100 runs were computed, as well as the K, CVM, and AD statistics (7-10), (7-11), and (7-12). The average NEES curves for the four batch lengths are plotted in figures 2 through 5. The K, CVM, and AD curves are plotted in figures 6 through 9.

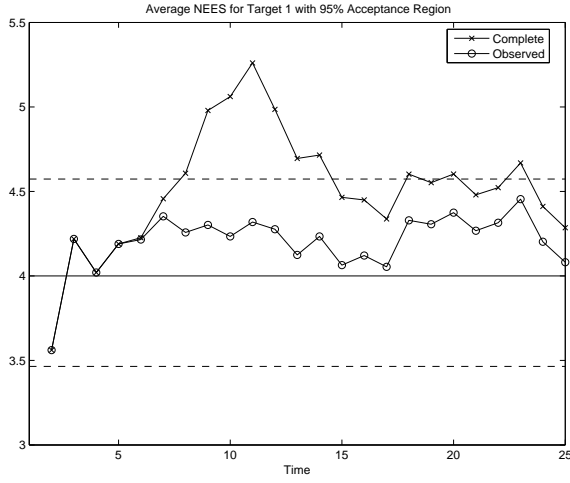
In figures 2 through 5, the horizontal solid line indicates the mean of the chi-square distribution with degrees of freedom  $q = 4$ ; the area between the horizontal dashed lines indicates the 95% acceptance region for the null hypothesis (7-5). It is clear from these plots that the posterior complete information matrix  $I_{\Theta|X}$  yields inconsistent estimates of estimation error in the vicinity of the crossing. Indeed, the average NEES curves associated with  $I_{\Theta|X}$  rise well above the acceptance regions near the crossing regardless of batch length, indicating overly optimistic estimates of estimation error; said another way, the error-covariance matrices computed using the posterior complete information matrix are too small in the vicinity of the crossing, where the measurement-to-source assignment uncertainty is large. In this region, the information lost to the missing data (measurement-to-source assignments) is significant, and the missing information term (the second term) in expression (5-75) for the posterior observed information matrix  $I_{\Theta|Y}$  is nonzero. From (5-75), it follows that the error-covariance matrix computed using the posterior observed information matrix  $I_{\Theta|Y}$  is always at least as large as the error-covariance matrix computed using the posterior complete information matrix  $I_{\Theta|X}$ . Hence, from (7-3) and (7-4), it follows that the average NEES curves computed using  $I_{\Theta|Y}$  in figures 2 through 5 are always bounded above by those computed using  $I_{\Theta|X}$ .

It is also clear from these figures that estimator consistency deteriorates with smaller batch length, in the sense that more average NEES values fall outside of the acceptance region as batch length decreases. This result is summarized in table 1, which records the percentages of the average NEES values for both targets and for all 25 sample times that fall within the 95% acceptance region. The shaded boxes in this table contain the percentages associated with the average NEES values computed using the posterior observed information matrix. The containment statistics for batch lengths of 25 and 10 indicate that the inverse of the posterior observed information matrix gives a consistent estimate of estimation error for this example; in both cases, 95.8% of the average NEES values fall within the 95% containment region. The containment statistics drop by 4.1 and 10.4 percentage points for batch lengths of 5 and 1, respectively. Interestingly, the containment statistics for the average NEES values computed using the posterior complete information matrix increase with decreasing batch length. These results are recorded in the unshaded boxes in table 1. The reason for this trend is not clear. In any event, these containment statistics are always worse than the corresponding statistics computed using the posterior observed information matrix, and all are well below the expected 95% containment level.

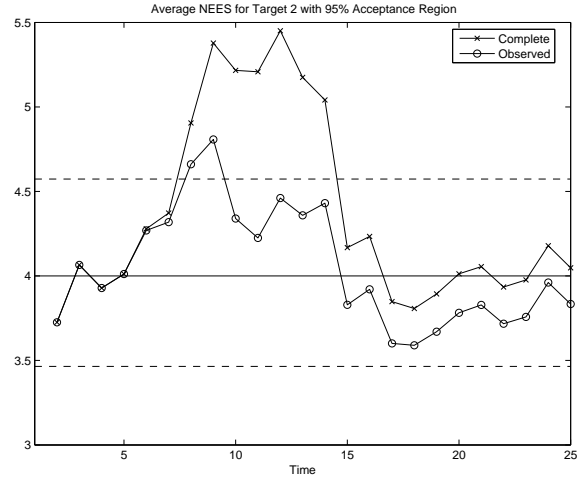
Figures 6 through 9 show plots of the K, CVM, and AD statistics as functions of sampling time for each of the four batch lengths. Recall that the test of estimator consistency based on these statistics is one-sided; the area below the horizontal dashed line in these plots is the 95% acceptance region for the test. All of the results discussed above for the average NEES curves hold for the K, CVM, and AD curves shown here, with one exception: the curves for the K, CVM, and AD statistics computed using the posterior observed information matrix  $I_{\Theta|Y}$  are not necessarily bounded above by those computed from the posterior complete information matrix  $I_{\Theta|X}$ . Nevertheless, the containment statistics in table 1 indicate that  $I_{\Theta|Y}^{-1}$  is a consistent estimate of estimation error, while  $I_{\Theta|X}^{-1}$  is not. Of the three statistics, the AD statistic is closest in behavior to the average NEES.



*Figure 1. Distance Between Targets in  $x$ -Dimension in Units of Measurement Standard Deviation for Crossing Targets Example*

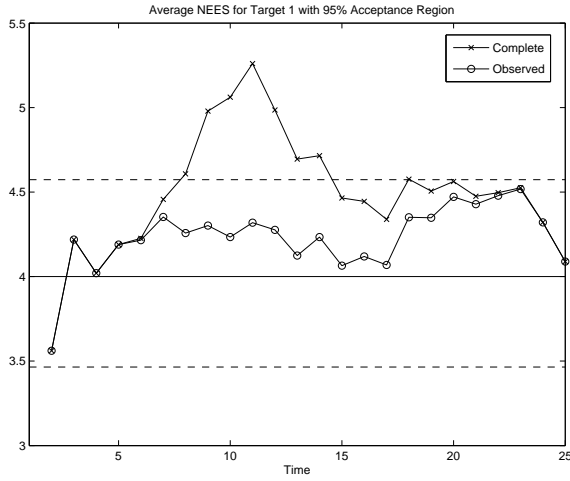


(a) Target 1.

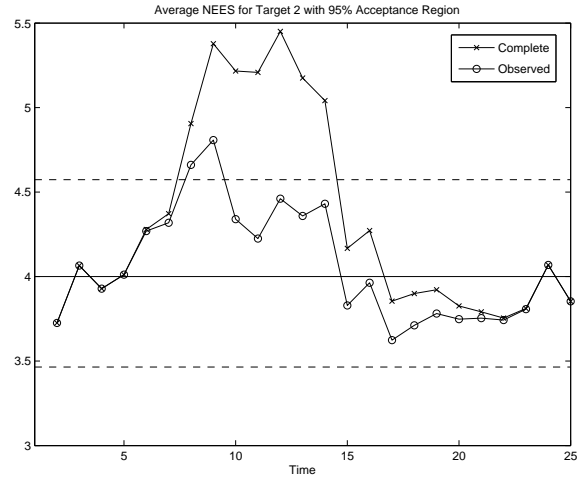


(b) Target 2.

**Figure 2. Average NEES with 95% Acceptance Region for Crossing Targets Example with Batch Length 25, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**

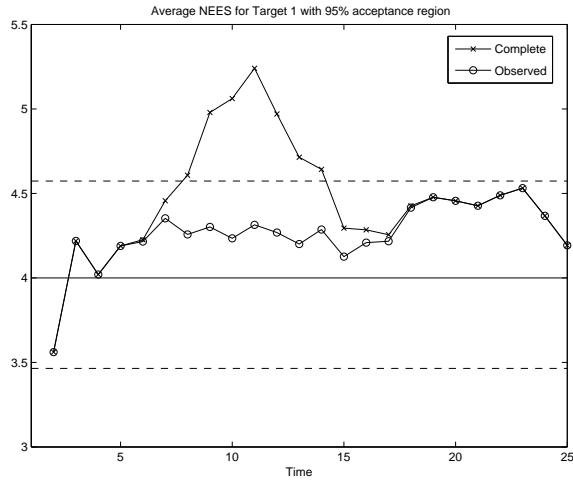


(a) Target 1.

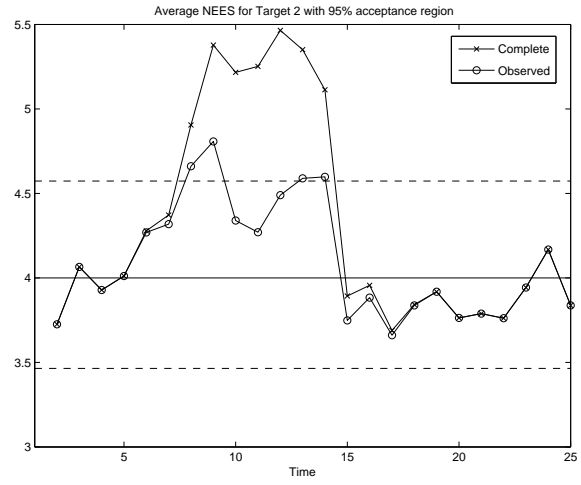


(b) Target 2.

**Figure 3. Average NEES with 95% Acceptance Region for Crossing Targets Example with Batch Length 10, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**

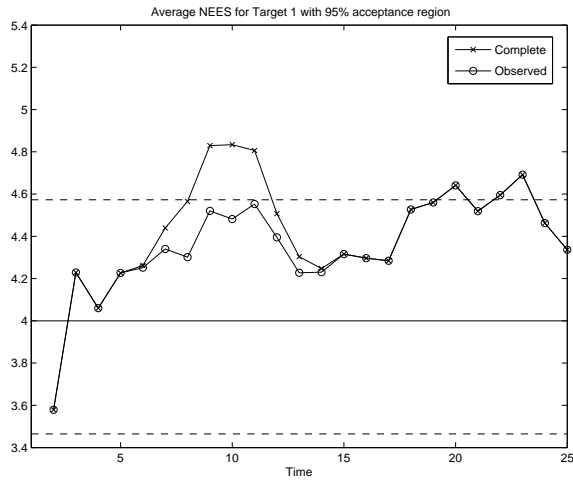


(a) Target 1.

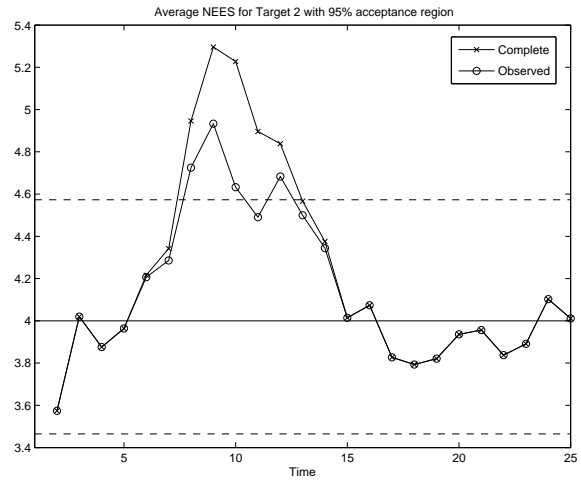


(b) Target 2.

**Figure 4. Average NEES with 95% Acceptance Region for Crossing Targets Example with Batch Length 5, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**

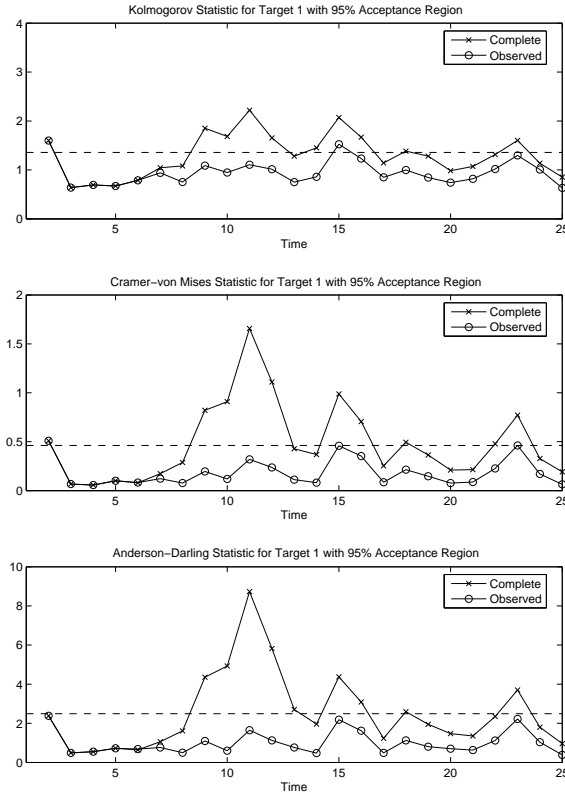


(a) Target 1.

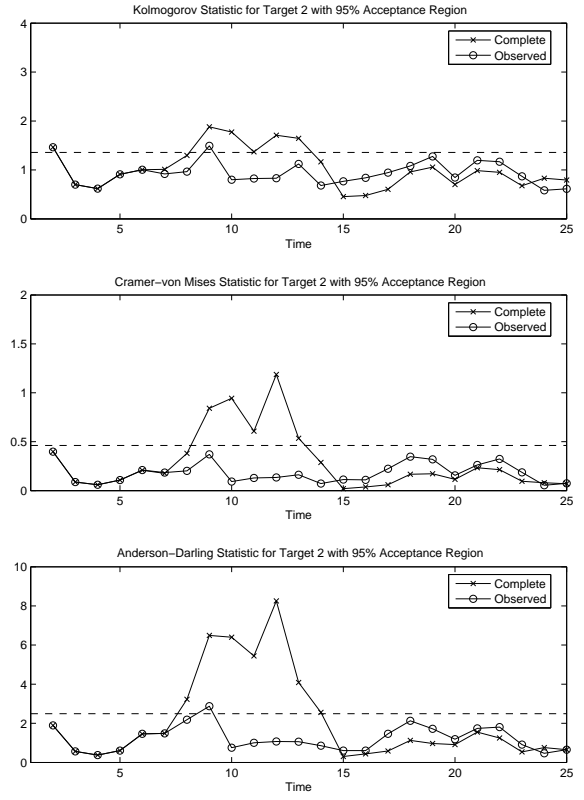


(b) Target 2.

**Figure 5. Average NEES with 95% Acceptance Region for Crossing Targets Example with Batch Length 1, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**

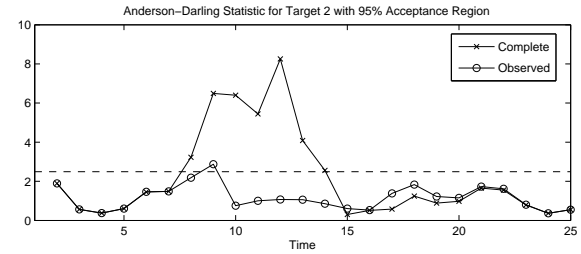
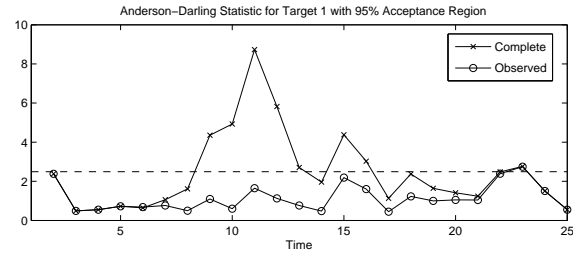
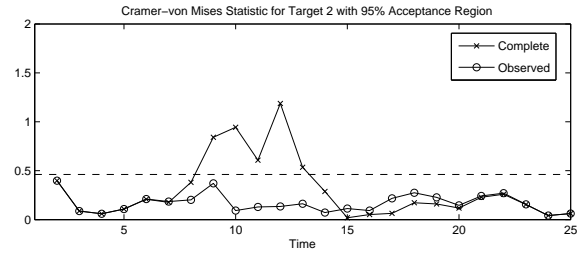
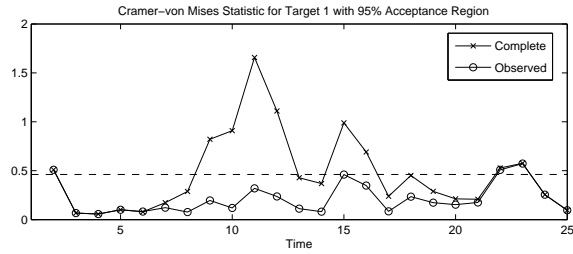
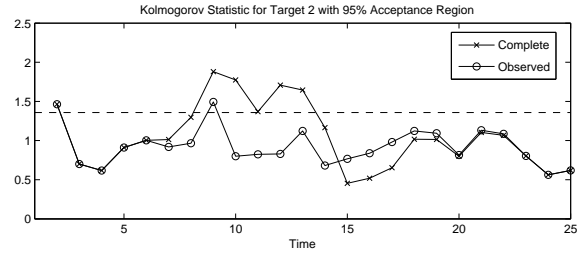
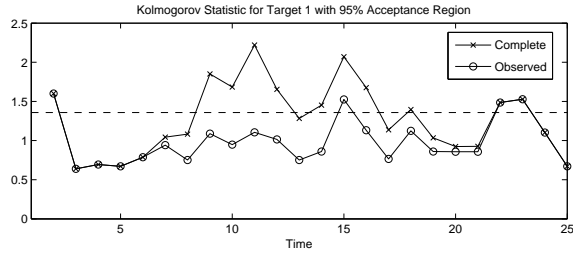


(a) Target 1.



(b) Target 2.

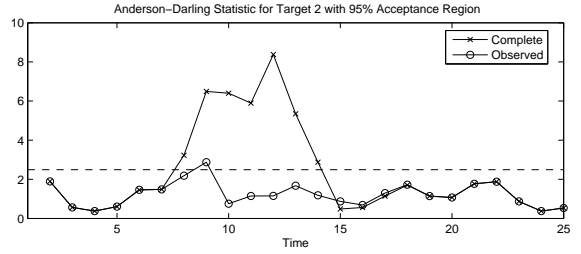
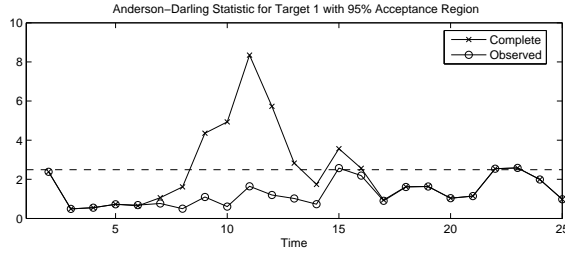
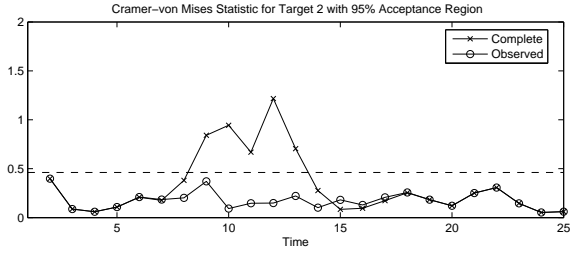
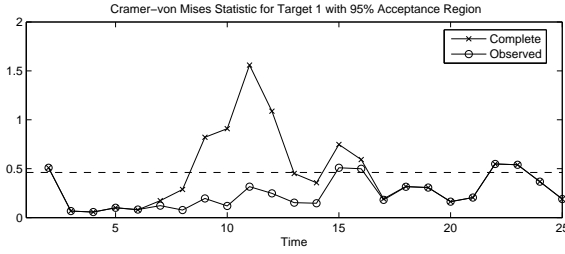
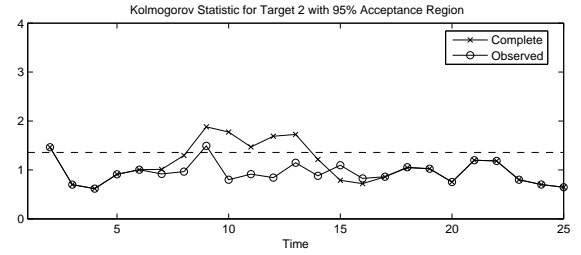
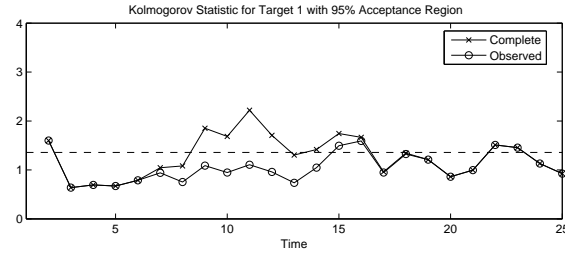
**Figure 6.  $K$ , CVM, and AD Statistics with 95% Acceptance Regions for Crossing Targets Example with Batch Length 25, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**



(a) Target 1.

(b) Target 2.

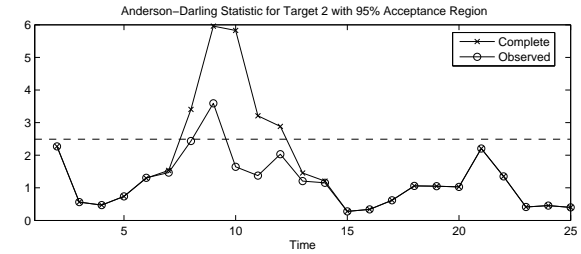
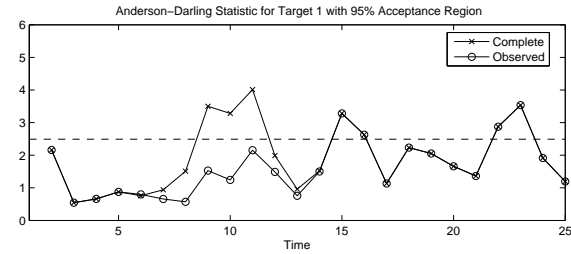
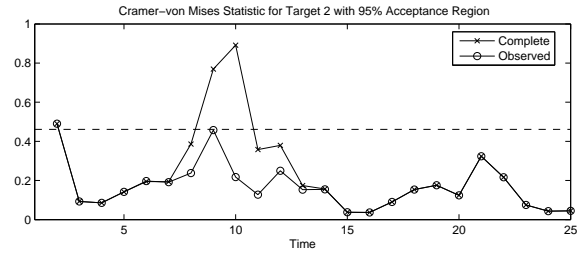
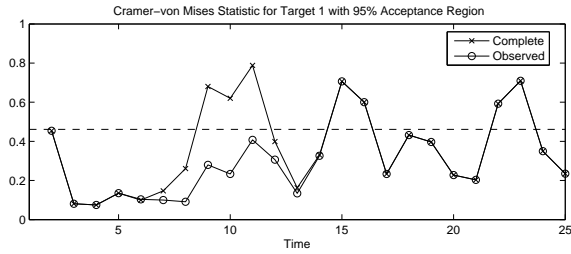
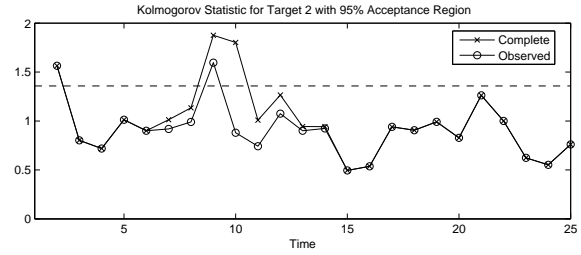
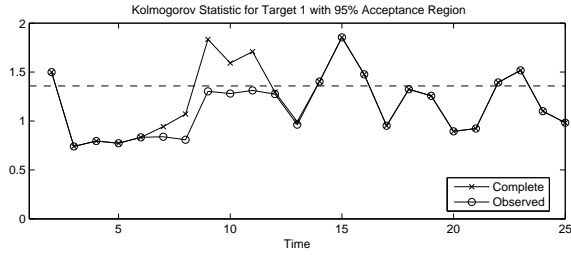
**Figure 7. *K*, CVM, and AD Statistics with 95% Acceptance Regions for Crossing Targets**  
**Example with Batch Length 10, Computed Using Posterior Complete Information Matrix**  
**(crosses) and Posterior Observed Information Matrix (circles)**



(a) Target 1.

(b) Target 2.

**Figure 8.  $K$ , CVM, and AD Statistics with 95% Acceptance Regions for Crossing Targets Example with Batch Length 5, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**



(a) Target 1.

(b) Target 2.

**Figure 9.  $K$ , CVM, and AD Statistics with 95% Acceptance Regions for Crossing Targets Example with Batch Length 1, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**



***Table 1. Percentage of NEES, K, CVM, and AD Values That Fall Within Their Respective 95% Acceptance Regions for the Crossing Targets Example (The first and second rows for each statistic correspond to use of the posterior complete and posterior observed information matrices, respectively, to compute the statistic.)***

Statistic	Batch Length			
	25	10	5	1
NEES	64.6	70.8	70.8	77.1
	95.8	95.8	91.7	85.4
K	68.8	64.6	66.7	75.0
	91.7	87.5	85.4	85.4
CVM	68.8	70.8	70.8	79.2
	97.9	91.7	89.6	89.6
AD	66.7	68.8	66.7	75.0
	97.9	95.8	91.7	89.6

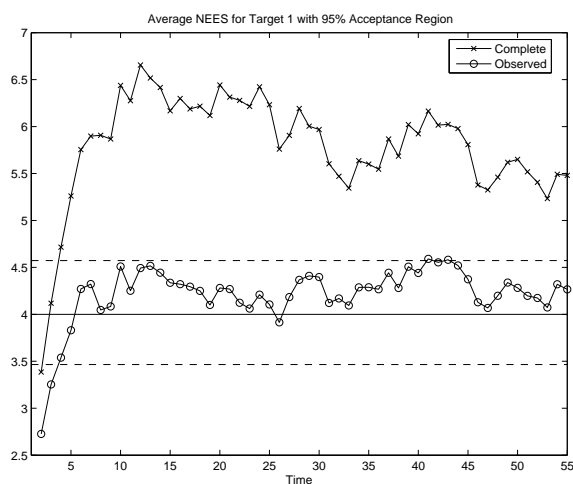
### 7.3 SINGLE TARGET IN CLUTTER

In this example, a single constant-velocity target travels in the  $xy$ -plane, with non-unity probability of detection  $P_D$  and non-zero probability of false alarm  $P_{FA}$ . In particular, a  $P_D$  of 0.9 is assumed fixed and known. Furthermore, it is assumed that at each sampling time  $t$ ,  $n_c(t)$  uniformly distributed clutter points are observed in a square coverage region centered at the true position of the target and with sides of length  $20r$ , where  $r = 0.075$  is the  $xy$ -position measurement standard deviation in each dimension from the previous example. The number of clutter points  $n_c(t)$  is assumed to be Poisson distributed with mean  $\lambda_c V = 4$ , where  $V$  is the volume of the coverage region (in this case  $1.5 \times 1.5 = 2.25$ ), and  $\lambda_c$  is the clutter density in this region ( $\lambda_c = 1.78$  in this case). Thus, on average, four uniformly distributed clutter points are expected in a  $1.5 \times 1.5$  region about the true target position; the probability of observing at least one clutter point in this region is  $\text{Prob}\{n_c(t) > 0\} = 0.98$ .

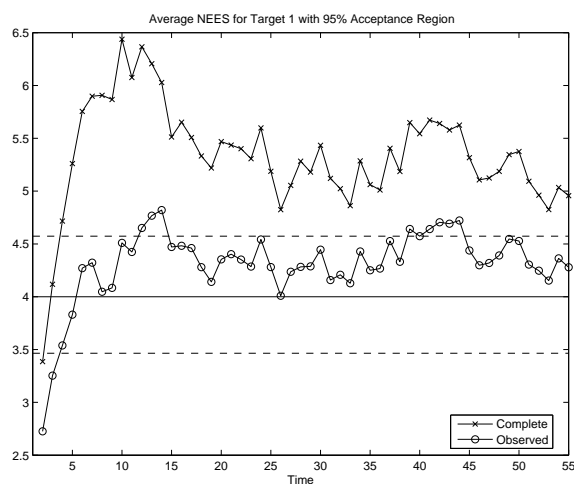
At time  $t = 0$ , the target is at  $xy$ -position  $(1, 0)$  with  $xy$ -velocity  $(0, 2)$ . At each time  $t = 1, \dots, 55$ , at most one measurement of target  $xy$ -position, and  $n_c(t)$  clutter points (false measurements) are obtained, each distributed as described above. The mean vector for the prior distribution of the target is taken to be the true position and velocity vector of the target at time  $t = 0$ , so that  $\eta_1 = (1, 0, 0, 2)$ . The prior covariance matrix for the target is taken to be  $1 \times 10^{-2}$  times the matrix (7-13), and the process noise covariance matrix is taken to be the matrix (7-14). The prior distribution for the target state at time  $t_0$  is made more informative (via the multiplicative factor  $1 \times 10^{-2}$ ) in this example to compensate for the well-known difficulty of initializing a tracker in clutter. There are various other ways to address this problem, but they are outside the scope of this report.

The PMHT model as described in section 5.2 must be modified to account for false alarms; that is, a clutter model must be added to account for observations that do not originate from a target. This is accomplished as described in Gauvrit et al. [9] by adding a uniform density function to the mixture density function for each observation. The impact of this clutter model on the update equations and information matrix computations for the linear Gauss-Markov mixture is primarily confined to the conditional measurement-to-source assignment probabilities. Additionally, some of the information matrix expressions (5-47) through (5-57) must change to reflect the addition of the clutter source to the measurement mixture model. These changes are listed in appendix C. Finally, for consistency with the assumption that at most one measurement originates from the target, the target mixing proportion  $\pi_1$  is set to 0.18, according to expression (C-12), which accounts for the probability of detection  $P_D$  and expected number of false alarms  $\lambda_c V$ . The clutter mixing proportion, denoted  $\pi_2$ , is then  $1 - \pi_1 = 0.82$ , and is held fixed for the simulation.

This simulation was run 100 times for each of the four batch lengths  $\bar{\rho} = 25, 10, 5$ , and 1. As for the crossing targets example, average NEES values and the K, CVM, and AD statistics were computed over these runs for these batch lengths using both the posterior complete information matrix and the posterior observed information matrix. The average NEES, K, CVM, and AD curves are plotted in figures 10 through 13. It is clear from these plots that, again, the posterior complete information matrix  $I_{\Theta|X}$  yields inconsistent estimates of estimation error. On the other hand, the posterior observed information matrix  $I_{\Theta|Y}$  gives consistent estimates in this example, at least for large enough batch length. These results are summarized in table 2, which records the percentages of average NEES, K, CVM, and AD values that fall within their respective 95% acceptance regions. As for the crossing targets example, the values in the unshaded boxes correspond to the statistics computed using the posterior complete information matrix; the values in the shaded boxes correspond to the statistics computed using the posterior observed information matrix. For the percentages in this table, only those statistics computed after sampling time  $t = 5$  were counted, since each of the statistics is initially skewed by the combination of informative prior information and good initialization. A randomized initialization scheme would perhaps have eliminated this trend, but the present scheme was deemed sufficient for this demonstration. In any event, the containment statistics for batch length 25 indicate that  $I_{\Theta|Y}^{-1}$  is a consistent estimate of the error-covariance matrix for this simulation. Again, as for the crossing targets example, the AD statistic is closest in behavior to the average NEES.

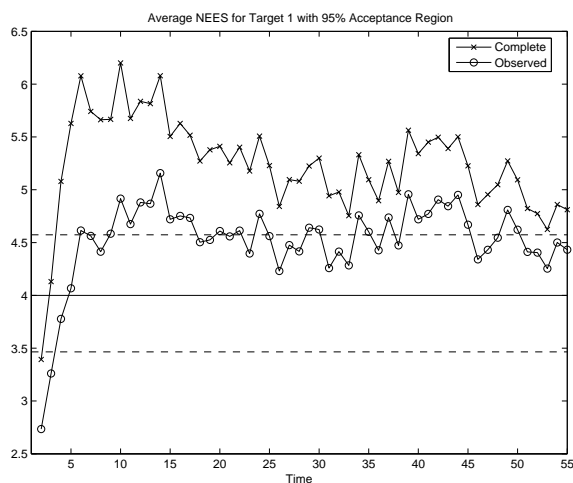


(a)  $\bar{\rho} = 25$ .

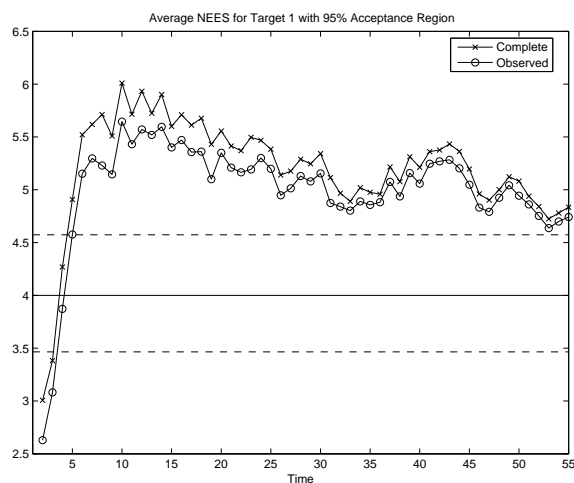


(b)  $\bar{\rho} = 10$ .

**Figure 10. Average NEES with 95% Acceptance Region for Single Target in Clutter**  
**Example with Batch Lengths 25 and 10, Computed Using Posterior Complete Information**  
**Matrix (crosses) and Posterior Observed Information Matrix (circles)**

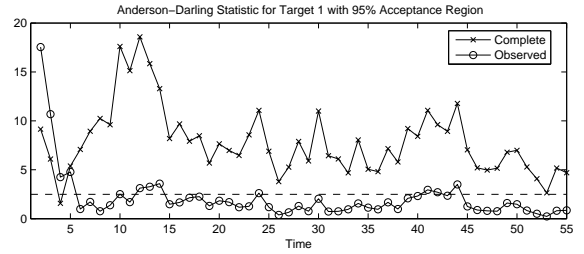
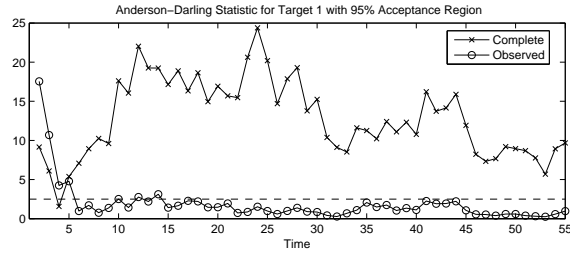
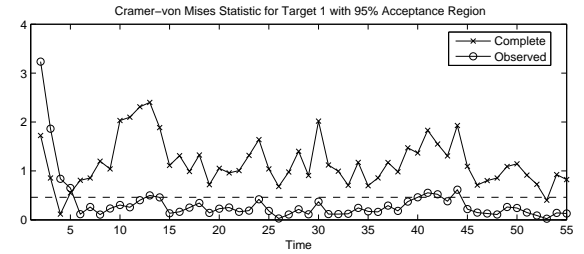
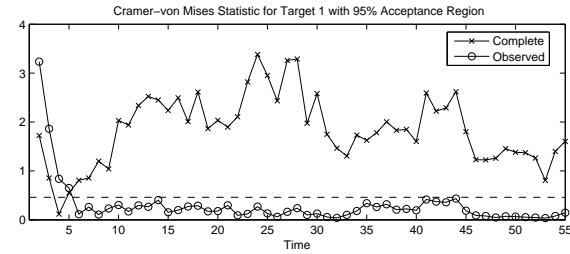
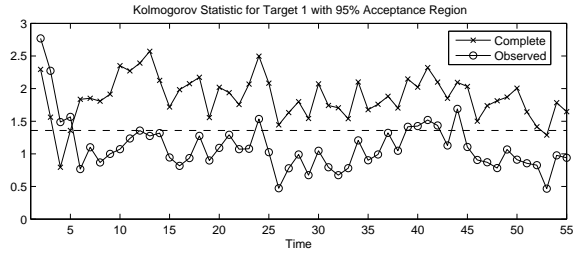
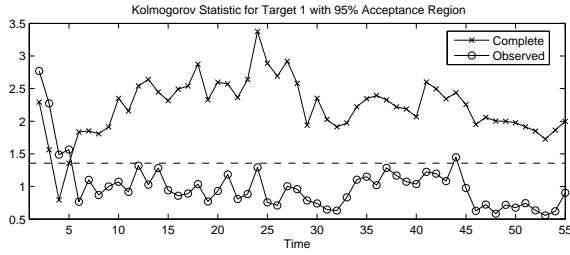


(a)  $\bar{\rho} = 5$ .



(b)  $\bar{\rho} = 1$ .

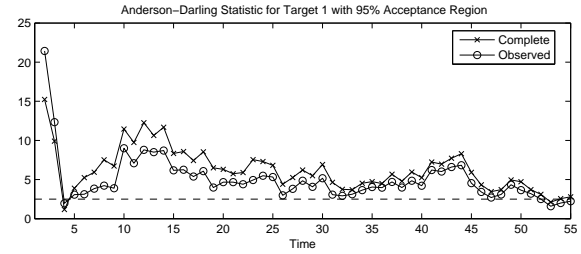
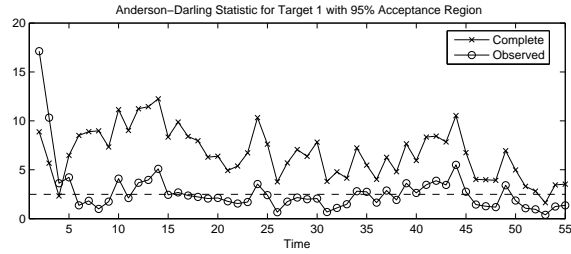
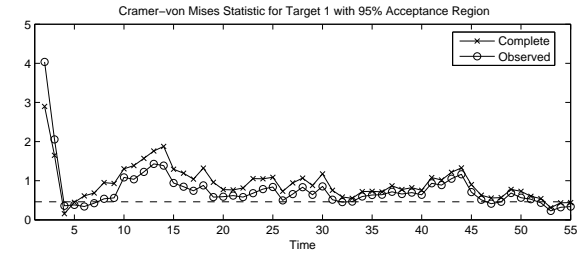
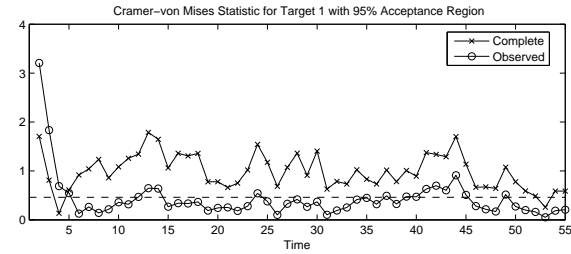
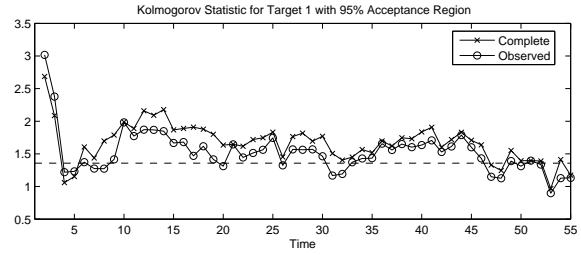
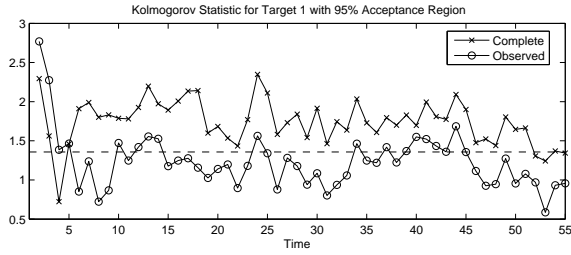
**Figure 11. Average NEES with 95% Acceptance Region for Single Target in Clutter**  
**Example with Batch Lengths 5 and 1, Computed Using Posterior Complete Information**  
**Matrix (crosses) and Posterior Observed Information Matrix (circles)**



(a)  $\bar{\rho} = 25$ .

(b)  $\bar{\rho} = 10$ .

**Figure 12. K, CVM, and AD Statistics with 95% Acceptance Region for Single Target in Clutter Example with Batch Lengths 25 and 10, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**



(a)  $\bar{\rho} = 5$ .

(b)  $\bar{\rho} = 1$ .

**Figure 13. *K*, CVM, and AD Statistics with 95% Acceptance Region for Single Target in Clutter Example with Batch Lengths 5 and 1, Computed Using Posterior Complete Information Matrix (crosses) and Posterior Observed Information Matrix (circles)**

**Table 2. Percentage of NEES, K, CVM, and AD Values That Fall Within Their Respective 95% Acceptance Regions for the Single Target in Clutter Example (The first and second rows for each statistic correspond to use of the posterior complete and posterior observed information matrices, respectively, to compute the statistic.)**

Statistic	Batch Length			
	25	10	5	1
NEES	0.0	0.0	0.0	0.0
	96.0	84.0	46.0	0.0
K	0.0	2.0	6.0	8.0
	98.0	88.0	74.0	26.0
CVM	0.0	2.0	2.0	6.0
	100.0	92.0	76.0	16.0
AD	0.0	0.0	2.0	4.0
	96.0	84.0	66.0	6.0





## 8. CONCLUSIONS

### 8.1 SUMMARY OF FINDINGS

An analytical approach for computing the observed information matrix for an important class of mixture models, called dynamic mixtures, is developed in this report. Dynamic mixtures are useful models for data originating from a number of distinct moving sources. Multiple target tracking is one application of these models; PMHT is the primary example of a dynamic mixture-based approach to multiple target tracking. In the basic PMHT model, a Gaussian mixture is used to describe the distribution of the measurements from each target, and a linear Gauss-Markov process model is used to describe the target dynamics.

An important finding of this report is the precise statistical interpretation of the error-covariance matrices for the PMHT track estimates in terms of the observed information matrix computations for these estimates. In particular, it is shown that the error-covariance matrices obtained from the Kalman smoothing filter for each target state sequence at the final EM iteration are the diagonal blocks of the inverse of the posterior complete information matrix for each sequence. Therefore, these error-covariance matrices provide only part of the information required to compute error-covariance matrices for the state estimates. In short, the error-covariance matrices obtained from the Kalman smoothing filters do not account for the information lost to the missing data, that is, the missing measurement-to-target assignments.

Another important finding of this report is the impact of measurement-to-source assignment uncertainty on estimator consistency. Specifically, for two common target tracking scenarios (two crossing targets, and a single target in clutter), it is shown that the posterior complete information matrix yields inconsistent estimates of estimation error when there is significant assignment uncertainty, while the posterior observed information matrix gives consistent estimates (for sufficient batch length). In each scenario, the standard chi-square test for the distribution of the average NEES is used to test for estimator consistency. Additionally, new tests for estimator consistency based on the EDF of the NEES are introduced; these tests are shown to produce results comparable to those of the standard NEES test.

### 8.2 ALTERNATIVE APPROACHES

While Louis's approach for computing the observed information matrix when using the EM method can be applied to any incomplete data problem, there are almost surely problems for which the required expressions, though based on complete data statistics, are difficult to derive analytically or compute numerically, or both. For these problems, the supplemented EM (SEM) method of Meng and Rubin [5] is an attractive alternative. Their method generalizes an observation made by Smith [44] in his discussion of Dempster et al. [1]. Based on

his analysis of the standard errors of a simple example in their paper, Smith alludes to the following general relationship between the observed data error-variance  $v_o$  and the complete data error-variance  $v_c$  of the maximum likelihood estimate for the scalar parameter  $\theta$ :

$$v_o = \frac{1}{1-r} v_c, \quad (8-1)$$

where, using the language of this report,  $v_o$  is the inverse of the observed information matrix (a scalar in this case),  $v_c$  is the inverse of the complete information matrix, and  $r$  is the rate of convergence of the EM method which, for large values of the iteration index  $k$ , is approximated by

$$r = \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)} - \theta^{(k-1)}}. \quad (8-2)$$

Thus, the observed data error-variance is obtained by inflating the complete data error-variance by the factor  $1/(1-r)$ . Meng and Rubin rewrite (8-1) in the statistically more appealing form

$$v_o = v_c + \Delta v, \quad (8-3)$$

where

$$\Delta v = \frac{r}{1-r} v_c \quad (8-4)$$

is interpreted as the increase in error-variance due to the missing data. Among the contributions of their paper are the analogous matrix version of (8-3), and computations for the matrix versions of  $v_c$  and  $r$ . Computation of the rate-of-convergence matrix  $r$  involves numerical differentiation of the implicit mapping  $\mathcal{M} : \Omega \rightarrow \Omega$  from the parameter space  $\Omega$  to itself defined by the EM method such that

$$\theta^{(k+1)} = \mathcal{M}(\theta^{(k)}) \quad \text{for } k = 0, 1, 2, \dots \quad (8-5)$$

However, unlike approaches such as Carlin's [45] that use numerical differentiation to obtain the error-covariance matrix directly from the observed data support function, SEM uses only numerical differentiation to obtain the increase due to the missing data to be added to the complete data error-covariance matrix. Hence, Meng and Rubin claim that SEM is typically more stable because the correction obtained by numerical differentiation is added to the complete data error-covariance matrix, which often can be obtained analytically and is usually the dominant term. Meng and Rubin do not include mixture models among the examples in their paper, although there is no impediment to using SEM for this problem. The use of SEM for dynamic mixture models, and a comparison of this approach with Louis's approach developed for these models in this report, are left as future work.

### 8.3 FUTURE INVESTIGATIONS

Several topics for future investigation have already been proposed. These include:

1. The application of the SEM method to dynamic mixture models, and a comparison of this method for computing the error-covariance matrix with Louis's approach.
2. An examination of the accuracy of the observed information matrix for dynamic mixtures as a function of sample size, and a comparison of the observed information matrix with the Fisher information matrix for these models.
3. The exploration of efficient methods for computing the inverse of the observed information matrix for dynamic mixture models, including suboptimal procedures for computing the error-covariance matrix for large problems.

Additionally, there are at least two more topics worth pursuing.

The other contribution of Louis's paper [3] is a method for accelerating convergence of the EM iterations using the observed information and complete information matrices. Specifically, Louis shows that the updated estimate for  $\theta$  at the  $k$ th EM iteration can be refined in place via the following step:

$$\theta_*^{(k)} = \theta^{(k)} + I_Y^{-1}(y; \theta^{(k)}) I_X(x; \theta^{(k)}) (\theta^{(k)} - \theta^{(k-1)}). \quad (8-6)$$

The refinement  $\theta_*^{(k)}$  is an improvement over the update  $\theta^{(k)}$  in the sense that the former is closer to  $\hat{\theta} = \theta^{(\infty)}$  than the latter.\* Application of this acceleration method to the dynamic mixture models presented in this report would appear to be straightforward.

Finally, it is proposed that the observed information matrix computations developed here be extended to dynamic mixture models for grouped and truncated data. In practice, data are often grouped into a finite number of observation cells either intentionally (for example, to simplify data collection) or unintentionally, perhaps due to limitations of the data collection process. Additionally, if the number of observations in an observations cell cannot be reported for any reason, the grouped data are said to be truncated. In any event, grouping and truncating samples introduces additional missing data into the estimation problem, namely, the sample locations within the observed cells, and the numbers of samples and their locations in the truncated cells. Consequently, the EM method is a natural approach to maximum likelihood estimation for these problems. This approach is treated by several authors, including Dempster et al. [1] and McLachlan and Jones [46]. The latter authors explicitly treat finite mixture models for grouped and truncated data.

---

\*There is a transposition error between the matrices  $I_X$  and  $I_Y^{-1}$  in this expression in Louis's paper [3, expression (5.3)]. The error is corrected by Meilijson in [35, expression (11)].

Recent work has been done on estimation for stochastic dynamic mixture models for grouped and truncated data. In particular, Luginbuhl [47] and Luginbuhl and Willett [48, 49] apply the PMHT model to a histogram representation of discrete time Fourier transform data to estimate the parameters of general frequency modulated signals in noise. Of particular interest to this discussion is a derivation in [47] of the Fisher information matrix for the parameters in a univariate Gaussian mixture approximation to a one-dimensional histogram. This result is important to this work, as it indicates the potential existence of a closed-form Cramér-Rao lower bound on estimation error for the stochastic dynamic mixture model, and the PMHT model in particular, for grouped and truncated data. Hence, this result should provide an opportunity to compare, in the spirit of Efron and Hinkley [4], the relative accuracy of the observed information matrix versus the Fisher information matrix for dynamic mixture models and, by extension, for multiple target tracking.

## APPENDIX A

### APPROXIMATION TO THE OBSERVED INFORMATION MATRIX

Use of the empirical Fisher information matrix (3-21) as an approximation to the observed information matrix for independent and identically distributed data is justified in the following sense. (The argument presented here is a detailed version of the argument in McLachlan and Krishnan [50]). Consider the information matrix (3-5) for independent and identically distributed observations:

$$I_Y(y; \theta) = - \sum_{i=1}^n \frac{\partial^2 \lambda_{Y_i}(y_i; \theta)}{\partial \theta \partial \theta^\top}, \quad (\text{A-1})$$

where the support functions  $\lambda_{Y_i}$  in this case are all the same function. Recalling that  $\lambda_{Y_i}(y_i; \theta) = \log f_{Y_i}(y_i; \theta)$  and manipulating the derivatives on the right-hand side of (A-1) yields

$$\begin{aligned} - \sum_{i=1}^n \frac{\partial^2 \lambda_{Y_i}(y_i; \theta)}{\partial \theta \partial \theta^\top} &= - \sum_{i=1}^n \frac{\partial}{\partial \theta} \left[ \frac{1}{f_{Y_i}(y_i; \theta)} \frac{\partial f_{Y_i}(y_i; \theta)}{\partial \theta^\top} \right] \\ &= \sum_{i=1}^n \left[ \frac{1}{f_{Y_i}^2(y_i; \theta)} \frac{\partial f_{Y_i}(y_i; \theta)}{\partial \theta} \frac{\partial f_{Y_i}(y_i; \theta)}{\partial \theta^\top} - \frac{1}{f_{Y_i}(y_i; \theta)} \frac{\partial^2 f_{Y_i}(y_i; \theta)}{\partial \theta \partial \theta^\top} \right] \\ &= \sum_{i=1}^n \frac{\partial \lambda_{Y_i}(y_i; \theta)}{\partial \theta} \frac{\partial \lambda_{Y_i}(y_i; \theta)}{\partial \theta^\top} - \sum_{i=1}^n \frac{1}{f_{Y_i}(y_i; \theta)} \frac{\partial^2 f_{Y_i}(y_i; \theta)}{\partial \theta \partial \theta^\top}. \end{aligned} \quad (\text{A-2})$$

Now, the expected value of  $I_Y(y; \theta)$  evaluated at  $\theta^*$ , the true value of  $\theta$ , is the Fisher information matrix  $I(\theta^*)$ . But the second term in the previous expression has zero expectation. Indeed,

$$\begin{aligned} E \left[ \sum_{i=1}^n \frac{1}{f_{Y_i}(Y_i; \theta)} \frac{\partial^2 f_{Y_i}(Y_i; \theta)}{\partial \theta \partial \theta^\top} \right] &= \sum_{i=1}^n \int_{\mathcal{Y}_i} \frac{\partial^2 f_{Y_i}(y_i; \theta)}{\partial \theta \partial \theta^\top} dy_i \\ &= \sum_{i=1}^n \frac{d^2}{d\theta d\theta^\top} \int_{\mathcal{Y}_i} f_{Y_i}(y_i; \theta) dy_i \\ &= 0, \end{aligned} \quad (\text{A-3})$$

where interchangeability of derivatives and integrals has been assumed. Therefore, insofar as  $\hat{\theta} \rightarrow \theta^*$  and  $I_Y(y; \hat{\theta}) \rightarrow I(\theta^*)$  as  $n \rightarrow \infty$ , it follows that for large sample sizes

$$I_Y(y; \hat{\theta}) = - \sum_{i=1}^n \frac{\partial^2 \lambda_{Y_i}(y_i; \theta)}{\partial \theta \partial \theta^\top} \bigg|_{\theta=\hat{\theta}} \approx \sum_{i=1}^n \frac{\partial \lambda_{Y_i}(y_i; \theta)}{\partial \theta} \frac{\partial \lambda_{Y_i}(y_i; \theta)}{\partial \theta^\top} \bigg|_{\theta=\hat{\theta}} = I_e(y; \hat{\theta}). \quad (\text{A-4})$$

See McLachlan and Krishnan for examples of this approximation, and Redner and Walker [34] and Meilijson [35] for uses of the empirical observed information matrix to accelerate convergence of the EM method.



## APPENDIX B

### INVERSE OF THE GAUSS-MARKOV PRIOR COVARIANCE MATRIX

The equality of the matrix  $I_{(prior)}$ , given by expression (5-44), and the inverse of the Gauss-Markov prior covariance matrix  $P$ , given by expression (5-65) and recursions (5-67) and (5-68), is established by the following theorem:

**Theorem.** *Let*

$$\tilde{Q}_k = \begin{cases} -\Gamma, & k = 0, \\ Q_{k-1}, & k = 1, \dots, t, \end{cases}$$

where  $\Gamma, Q_0, Q_1, \dots, Q_{t-1}$  are positive-definite matrices, let

$$\Gamma_k = \begin{cases} \Gamma, & k = 0, \\ F_{k,k-1}\Gamma_{k-1}F_{k,k-1}^\top + Q_{k-1}, & k = 1, \dots, t, \end{cases}$$

and let

$$\Upsilon_k = F_{k+1,k}^\top \tilde{Q}_{k+1}^{-1} F_{k+1,k}, \quad k = 0, \dots, t-1.$$

Furthermore, let

$$P = \begin{bmatrix} \Gamma_0 & \Gamma_0 F_{10}^\top & \Gamma_0 F_{10}^\top F_{21}^\top & \cdots & \Gamma_0 F_{10}^\top \cdots F_{t,t-1}^\top \\ F_{10}\Gamma_0 & \Gamma_1 & \Gamma_1 F_{21}^\top & \cdots & \Gamma_1 F_{21}^\top \cdots F_{t,t-1}^\top \\ F_{21}F_{10}\Gamma_0 & F_{21}\Gamma_1 & \Gamma_2 & \cdots & \Gamma_2 F_{32}^\top \cdots F_{t,t-1}^\top \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F_{t,t-1} \cdots F_{10}\Gamma_0 & F_{t,t-1} \cdots F_{21}\Gamma_1 & F_{t,t-1} \cdots F_{32}\Gamma_2 & \cdots & \Gamma_t \end{bmatrix},$$

and let

$$J = \begin{bmatrix} -\tilde{Q}_0^{-1} + \Upsilon_0 & -F_{10}^\top \tilde{Q}_1^{-1} & 0 & \cdots & 0 \\ -\tilde{Q}_1^{-1} F_{10} & \tilde{Q}_1^{-1} + \Upsilon_1 & -F_{21}^\top \tilde{Q}_2^{-1} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\tilde{Q}_{t-1}^{-1} F_{t-1,t-2} & \tilde{Q}_{t-1}^{-1} + \Upsilon_{t-1} & -F_{t,t-1}^\top \tilde{Q}_t^{-1} \\ 0 & \cdots & 0 & -\tilde{Q}_t^{-1} F_{t,t-1} & \tilde{Q}_t^{-1} \end{bmatrix}.$$

Then, for each positive integer  $t$ , the matrix equality  $P^{-1} = J$  holds.

**Proof.** For a given positive integer  $t$ , use the Gauss-Jordan method to reduce the concatenated matrix  $[P, I]$  to the matrix  $[I, P^{-1}]$ , where  $I$  is the compatibly sized identity matrix. In particular, let  $G^{(l)}$  denote the result of row reduction after the  $l$ th step, so that  $G^{(0)} \equiv [P, I]$  and  $G^{(L)} \equiv [I, P^{-1}]$  for some  $L > 0$ . Moreover, let  $r_k^{(l)}$  denote the  $k$ th row of  $G^{(l)}$ . Then, the

reduction of the rectangular matrix  $[P, I]$  to the matrix  $[I, P^{-1}]$  terminates in three steps, as given by the following row recursions:

$$\begin{aligned} r_k^{(1)} &= \begin{cases} r_k^{(0)}, & k = 0, \\ F_{k,k-1}r_{k-1}^{(0)} - r_k^{(0)}, & k = 1, \dots, t, \end{cases} \\ r_k^{(2)} &= \begin{cases} r_k^{(1)} - \tilde{Q}_k F_{k+1,k}^\top \tilde{Q}_{k+1}^{-1} r_{k+1}^{(1)}, & k = 0, \dots, t-1, \\ r_k^{(1)}, & k = t, \end{cases} \\ r_k^{(3)} &= -\tilde{Q}_k^{-1} r_k^{(2)}, \quad k = 0, \dots, t. \end{aligned}$$

The recursions  $r_k^{(1)}$  and  $r_k^{(2)}$  produce all zeros below and above the diagonals of the left-half partitions of  $G^{(1)}$  and  $G^{(2)}$ , respectively; finally, the recursions  $r_k^{(3)}$  produce ones along the diagonal of the left-half partition of  $G^{(3)}$ , leaving  $G^{(3)} = [I, P^{-1}]$ . Inspection of the right-half partition of this matrix reveals the identity  $P^{-1} = J$ .



## APPENDIX C

### ADDING A CLUTTER MODEL TO PMHT

The changes required to add a clutter distribution to the PMHT model discussed in section 5.2 are presented in this appendix.

Let  $D_t$  denote the sensor coverage region at sampling time  $t$ , and let  $V(D_t)$  denote the volume of this region. In the example of section 7.3, the coverage region  $D_t$  is the square centered at the true position of the target at time  $t$  and with sides of length  $20r$ , where  $r$  is the  $xy$ -position measurement standard deviation, so that  $V(D_t) = 400r^2$ . Let  $u(s; G)$  denote the uniform density function with support  $G$ , so that

$$u(s; G) = \begin{cases} \frac{1}{V(G)}, & \text{if } s \in G, \\ 0, & \text{otherwise,} \end{cases} \quad (\text{C-1})$$

and let  $\pi_{m+1}$  denote the mixing proportion associated with the clutter source. Then, with the inclusion of this clutter model, the observed data likelihood function (5-36) for the linear Gauss-Markov dynamic mixture becomes

$$f_{Y|\Theta}(y|\theta) = \prod_{t=1}^T \prod_{i=1}^{n_t} \sum_{j=1}^{m+1} \pi_j f_j(y_{ti}|\theta), \quad (\text{C-2})$$

where

$$f_j(y_{ti}|\theta) = \begin{cases} \phi(y_{ti}|M_{jt}\mu_{jt}, R_{jt}), & j = 1, \dots, m, \\ u(y_{ti}; D_t), & j = m+1. \end{cases} \quad (\text{C-3})$$

Also, the constraint (4-13) must be expanded to include the mixing proportion  $\pi_{m+1}$ . The impact of this clutter model on the update equations and information matrix computations for the linear Gauss-Markov mixture is for the most part confined to the conditional measurement-to-source probabilities (5-39). These probabilities become, at the  $k$ th EM iteration,

$$w_{jti}^{(k)} = \frac{\pi_j^{(k)} f_j(y_{ti}|\theta^{(k)})}{\sum_{l=1}^{m+1} \pi_l^{(k)} f_l(y_{ti}|\theta^{(k)})}, \quad (\text{C-4})$$

for  $j = 1, \dots, m+1$ . Additionally, some of the information matrix computations (5-47) through (5-57) must change to reflect the addition of the clutter source to the mixture model for the measurements. These changes are as follows: expression (5-47) becomes

$$\langle S_{ti} \rangle_{\pi_j} = w_{jti}/\pi_j - w_{m+1,ti}/\pi_{m+1}, \quad j = 1, \dots, m; \quad (\text{C-5})$$

expressions (5-49) and (5-51) become

$$\langle B_{ti} \rangle_{\pi_j \pi_l} = \begin{cases} w_{jti}/\pi_j^2 + w_{m+1,ti}/\pi_{m+1}^2, & j = l, \\ w_{m+1,ti}/\pi_{m+1}^2, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (\text{C-6})$$

and

$$\langle B_{ti} \rangle_{\pi_j \mu_l} = 0, \quad j, l = 1 \dots, m, \quad (\text{C-7})$$

respectively; expressions (5-52) and (5-54) become, respectively,

$$\langle S_{ti} S_{ti}^T \rangle_{\pi_j \pi_l} = \begin{cases} w_{jti}/\pi_j^2 + w_{m+1,ti}/\pi_{m+1}^2, & j = l, \\ w_{m+1,ti}/\pi_{m+1}^2, & j \neq l, \end{cases} \quad j, l = 1, \dots, m, \quad (\text{C-8})$$

and

$$\langle S_{ti} S_{ti}^T \rangle_{\pi_j \mu_l} = e_t^\circ \otimes \frac{w_{jti}}{\pi_j} (y_{ti} - M_{jt} \mu_{jt})^T R_{jt}^{-1} M_{jt}, \quad j, l = 1, \dots, m; \quad (\text{C-9})$$

finally, expressions (5-55) and (5-57) become,

$$-\nabla_{\pi_j} \{ \nabla_{\pi_l} \lambda_\Theta(\theta) \}^T = 0, \quad j, l = 1, \dots, m, \quad (\text{C-10})$$

$$-\nabla_{\pi_j} \{ \nabla_{\mu_l} \lambda_\Theta(\theta) \}^T = 0, \quad j, l = 1, \dots, m. \quad (\text{C-11})$$

To be consistent with the standard tracking assumption that at most one observation at each sampling time is associated with each target, the following heuristic is used to set the target mixing proportions given fixed probability of detection  $P_D$ , clutter density  $\lambda_c$ , and sensor coverage region volume  $V$  (assumed here to be constant):

$$\pi_j = \frac{P_D}{m + \lambda_c V}, \quad j = 1, \dots, m. \quad (\text{C-12})$$

The clutter mixing proportion is then  $\pi_{m+1} = 1 - \pi_1 - \dots - \pi_m$ . This heuristic is slightly different than the one proposed by Rago et al. [38] and Willett et al. [39]. In particular, the denominator in their expression is a function of the number  $n_t$  of observations at time  $t$ . In either case, experience indicates that PMHT algorithm performance is relatively insensitive to precise values of the mixing proportions, and that rough approximations such as (C-12) are usually adequate.

## REFERENCES

1. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)," *Journal of the Royal Statistical Society B*, vol. 39, 1977, pp. 1–38.
2. G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.
3. T. A. Louis, "Finding the Observed Information Matrix when Using the EM Algorithm," *Journal of the Royal Statistical Society B*, vol. 44, 1982, pp. 226–233.
4. B. Efron and D. V. Hinkley, "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed versus Expected Fisher Information (with Discussion)," *Biometrika*, vol. 65, 1978, pp. 457–487.
5. X. L. Meng and D. B. Rubin, "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, vol. 86, 1991, pp. 899–909.
6. P. J. Green, "On Use of the EM Algorithm for Penalized Likelihood Estimation," *Journal of the Royal Statistical Society B*, vol. 52, no. 3, 1990, pp. 443–452.
7. M. R. Segal, P. Bacchetti, and N. P. Jewell, "Variances for Maximum Penalized Likelihood Estimates Obtained via the EM Algorithm," *Journal of the Royal Statistical Society B*, vol. 56, 1994, pp. 345–352.
8. R. L. Streit and T. E. Luginbuhl, "Probabilistic Multi-Hypothesis Tracking," Technical Report 10,428, Naval Undersea Warfare Center Division, Newport, RI, 15 February 1995.
9. H. Gauvrit, J.-P. Le Cadre, and C. Jauffret, "A Formulation of Multitarget Tracking as an Incomplete Data Problem," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 4, October 1997, pp. 1242–1257.
10. D. B. Reid, "An Algorithm for Tracking Multiple Targets," *IEEE Transactions on Automatic Control*, vol. AC-24, no. 6, December 1979, pp. 843–854.
11. R. L. Streit, "PMHT Algorithms for Multi-Frame Assignment," in *Proceedings of the 9th International Conference on Information Fusion*, Florence, Italy, International Society of Information Fusion (<http://www.isif.org/>), to appear.
12. D. Avitzour, "A Maximum Likelihood Approach to Data Association," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 28, no. 2, 1992, pp. 560–566.

13. K. J. Molnar and J. W. Modestino, "Application of the EM Algorithm for the Multitarget/Multisensor Tracking Problem," *IEEE Transactions on Signal Processing*, vol. 46, no. 1, January 1998, pp. 115–129.
14. T. E. Fortmann, Y. Bar-Shalom, M. Scheffe, and S. B. Gelfand, "Detection Thresholds for Tracking in Clutter—A Connection Between Estimation and Signal Processing," *IEEE Transactions on Automatic Control*, vol. AC-30, no. 3, March 1985, pp. 221–229.
15. Y. Bar-Shalom and E. Tse, "Tracking in a Cluttered Environment with Probabilistic Data Association," *Automatica*, vol. 11, September 1975, pp. 451–460.
16. S. B. Gelfand, T. E. Fortmann, and Y. Bar-Shalom, "Adaptive Detection Threshold Optimization for Tracking in Clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 2, April 1996.
17. F. E. Daum, "Bounds on Performance of Multiple Target Tracking," *IEEE Transactions on Automatic Control*, vol. 35, no. 4, April 1990, pp. 443–445.
18. C. Jauffret and Y. Bar-Shalom, "Track Formation with Bearing and Frequency Measurements in Clutter," *IEEE Transactions on Signal Processing*, vol. 26, no. 6, November 2001, pp. 999–1010.
19. T. Kirubarajan and Y. Bar-Shalom, "Low Observable Target Motion Analysis Using Amplitude Information," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 4, October 1996.
20. P. Willett and Y. Bar-Shalom, "On the CRLB when Measurements are of Uncertain Origin," in *Proceedings of the 37th IEEE Conference on Decision & Control*, Tampa, FL, December 1998, pp. 743–747.
21. R. Niu, P. Willett, and Y. Bar-Shalom, "Matrix CRLB Scaling Due to Measurements of Uncertain Origin," *IEEE Transactions on Signal Processing*, vol. 49, no. 7, July 2001, pp. 1325–1335.
22. L. I. Perlovsky, "Cramér-Rao Bound for Tracking in Clutter and Tracking Multiple Objects," *Pattern Recognition Letters*, vol. 18, 1997, pp. 283–288.
23. L. I. Perlovsky, "Cramér-Rao Bounds for the Estimation of Normal Mixtures," *Pattern Recognition Letters*, vol. 10, 1989, pp. 141–148.
24. J.-P. Le Cadre, H. Gauvrit, and F. Trarieux, "Approximations of the Cramér-Rao Bound for Multiple-Target Motion Analysis," *IEE Proceedings - Radar, Sonar, and Navigation*, vol. 147, no. 3, June 2000, pp. 105–113.

25. C. Hue, J-P. Le Cadre, and P. Pérez, "Posterior Cramér-Rao Bounds for Multi-Target Tracking," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 1, January 2006.
26. M. L. Graham and R. L. Streit, "The Cramér-Rao Bound for Multiple Target Tracking Algorithms," Technical Report 10,406, Naval Undersea Warfare Center Division, Newport, RI, 30 September 1994.
27. J. Cai, A. Sinha, and T. Kirubarajan, "EM-ML Algorithm for Track Initialization Using Possibly Noninformative Data," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 3, July 2005, pp. 1030–1048.
28. C. F. J. Wu, "On the Convergence Properties of the EM algorithm," *Annals of Statistics*, vol. 11, 1983, pp. 95–103.
29. H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1946.
30. G. Casella and R. L. Berger, *Statistical Inference*, Duxbury Press, Belmont, CA, 1990.
31. R. A. Redner, R. J. Hathaway, and J. C Bezdek, "Estimating the Parameters of Mixture Models with Modal Estimators," *Communications in Statistics: Theory and Methods*, vol. 16, no. 9, 1987, pp. 2639–2660.
32. A. W. F. Edwards, *Likelihood*, Expanded edition, The Johns Hopkins University Press, Baltimore, MD, 1992.
33. T. Orchard and M. A. Woodbury, "A Missing Information Principle: Theory and Applications," in *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, University of California Press, Berkeley, CA, 1972, pp. 697–715.
34. R. A. Redner and H. F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, vol. 26, no. 2, 1984, pp. 195–239.
35. I. Meilijson, "A Fast Improvement to the EM Algorithm on its Own Terms," *Journal of the Royal Statistical Society B*, vol. 51, no. 1, 1982, pp. 127–138.
36. J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*, Prentice Hall PTR, Englewood Cliffs, NJ, 1995.
37. G. J. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, New York, 2000.
38. C. Rago, P. Willett, and R. L. Streit, "Direct Data Fusion Using the PMHT," in *Proceedings of the 1995 American Control Conference*, Seattle, WA, 21–23 June 1995, pp. 1678–1682.

39. P. Willett, Y. Ruan, and R. L. Streit, "PMHT: Problems and Some Solutions," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 3, July 2002, pp. 738–754.
40. Y. Bar-Shalom and X. R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*, YBS Publishing, Storrs, CT, 1995.
41. R. B. D'Agostino and M. A. Stephens, Eds., *Goodness-of-Fit Techniques*, Marcel Dekker, New York, 1986.
42. A. Stuart, J. K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, vol. 2A: *Classical Inference and the Linear Model*, 6th edition, Arnold Publishers, London, 1999.
43. A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, vol. 1: *Distribution Theory*, 6th edition, Arnold Publishers, London, 1994.
44. C. A. B. Smith, "Discussion of 'Maximum Likelihood from Incomplete Data via the EM Algorithm,' by A. P. Dempster, N. M. Laird, and D. B. Rubin," *Journal of the Royal Statistical Society B*, vol. 39, 1977, pp. 1–38.
45. J. B. Carlin, *Seasonal Analysis of Economic Time Series*, Ph.D. thesis, Harvard University, Department of Statistics, Cambridge, MA, 1987.
46. G. J. McLachlan and P. N. Jones, "Fitting Mixture Models to Grouped and Truncated Data via the EM Algorithm," *Biometrics*, vol. 44, 1988, pp. 571–578.
47. T. E. Luginbuhl, *Estimation of General Discrete-Time FM Processes*, Ph.D. thesis, University of Connecticut, Electrical Engineering Department, Storrs, CT, 1999.
48. T. E. Luginbuhl and P. Willett, "Tracking a General, Frequency Modulated Signal in Noise," in *Proceedings of the 38th IEEE Conference on Decision & Control*, Phoenix, AZ, December 1999, pp. 5076–5081.
49. T. E. Luginbuhl and P. Willett, "Estimating the Parameters of General Frequency Modulated Signals," *IEEE Transactions on Signal Processing*, vol. 52, 2004, pp. 117–131.
50. G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley & Sons, New York, 1997.

## INITIAL DISTRIBUTION LIST

**Addressee**

**No. of Copies**

Defense Technical Information Center

1 (CD-ROM)

